CYUT Chinese Textual Entailment Recognition System for NTCIR-10 RITE-2

Shih-Hung Wu, Shan-Shan Yang
CSIE, Chaoyang University of Technology,
Taiwan, R.O.C
{shwu, s10027619}@cyut.edu.tw

Liang-Pu Chen, Hung-Sheng Chiu, Ren-Dar Yang
Institute for Information Industry, Taipei, Taiwan, R.O.C
{eit, bbchiu, rdyang}@iii.org.tw

ABSTRACT
Textual Entailment (TE) is a critical issue in natural language processing (NLP). In this paper we report our approach to the Chinese textual entailment and the system result on NTCIR-10 RITE-2 both simplified and traditional Chinese dataset. Our approach is based on more observation on training data and finding more types of linguistic features. The approach is a complement to the traditional machine learning approach, which treat every pair in a standard process. In the official runs, we tested three types of entailment features, i.e. the usage of negative words, time expression, and numbers. The experimental result is promising; we find this extensible approach can include more types.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Natural language understanding, Textual Entailment

General Terms
Experimentation

Keywords
Chinese Textual Entailment, linguistic feature, classifier

Team Name
CYUT

Subtasks/Languages
RITE, RITE4QA, Simplified Chinese, Traditional Chinese

External Resources Used
LIBSVM, Stanford Parser

1. INTRODUCTION
TE can be briefly defined as :"Given a pair of sentences (t1, t2), a program has to decide whether the information in t2 can be inferred by t1." This is critical task in NLP recently. TE can be used in various applications, such as question answering system, information extraction, information retrieval, and machine translation[2][3]. Basic approach to TE is based on the semantic and syntactic similarity of the words in the sentences [4]. Once a system can decide whether t1 entail t2 or not, it can help a lot on finding useful information for the users as an information filter.

Most of the TE research focus on English, there are very few data and result in other languages such as Chinese or Japanese [5]. The RITE-1 shared task in NTCIR-9 in 2010 is the first event that proves Chinese dataset. As a continues event, RITE-2 shared task in NTCIR-10 also provides Chinese dataset [6]. The data set provide binary class textual entailment and multi-class textual entailment. Table 1 shows the examples of sentence pairs in the training corpus. Where forward type TE means the information in t2 can be inferred by t1, but the information in t1 cannot be inferred by t2. The bidirectional type TE means the information in the two sentences t1 and t2 are almost the same; one can be inferred by the other one. The contradiction type means the information in the two sentences cannot be true at the same time; they have different information and contradict to each other. The independent type means the pair does not belong to the previous three types; the information in t1 and t2 is irrelevant, one cannot entail or contradict to the other.

The goal of attending the task is to build a system based on available resources and test how the performance of the system. Our previous system is built based on a SVM classifier, which involved 12 textual features. By observation on the training corpus, we find that there are many different types of feature and should be treated separately. We tested in the official run three cases: i.e. the usage of negative words, time expression, and numbers. The experimental result is promising; we find this extensible approach can include more types.

The rest of the paper is organized as follows: we describe our methodology in section 2, and give the details of our system in section 3. The experimental result is shown in section 4. Final section is the conclusions and future works.

Table 1. Examples of five entailment relations

Type	Example
Forward	t1：田中耕一在製藥的發展上，對早期的藥物研發程序已產生了革命性的改變，加速新藥開發。 (The pharmaceutical development achievements of Koichi Tanaka have produced a revolutionary change on the early drug development process, and accelerate the development of new drugs.)
	t2：田中耕一其成就可加速新一代藥物的研發。 (The achievements of Koichi Tanaka can accelerate the research and development of a new generation of drugs.)
Bidirectional	t1：楊振寧、金庸、聖嚴及劉兆玄等人，以

	主題為「歲月的智慧—大師真情」展開對談， (Chen Ning Yang, Jin Yong, Sheng Yen and Liu Chao-hsuan, dialogue at the theme of "the wisdom of the years - the true feelings of masters")
	t2：這一場別開生面的「歲月的智慧」大師真情對談，有楊振寧、金庸、聖嚴法師及劉兆玄四人。 (In this special "wisdom of the years" masters talk with true feelings, Chen Ning Yang, Jin Yong, Master Sheng Yen and Liu Chao-hsuan four persons attended.)
Contradiction	t1：2004年奧運羅馬尼亞籍世界冠軍拳手西蒙在16強賽以24：36點數落後、敗給埃及的亞瑟。 (2004 Olympic Romanian world champion boxer Simon lost to Arthur from Egypt, with 24:36 points behind in the round of final 16.)
	t2：2004 年奧運羅馬尼亞籍世界冠軍拳手西蒙在 16 強賽以 24：36 點數領先、大敗埃及的亞瑟。 (2004 Olympic Romanian world champion boxer Simon defeated Egyptian Arthur, in the Round of final 16 with 24:36 points lead.)
Independence	t1：印尼發生之最大地震，掀起 10 公尺高牆般的巨浪海嘯。 (The largest earthquake in Indonesia sets off waves of tsunami 10 meters high.)
	t2：地震發生後，明打威群島一些島嶼遭到3米多高的巨浪衝擊，島上數百所房屋被卷走。 (After the earthquake, some islands of the Mentawai Islands have been struck by a 3-meter-high waves, hundreds of houses on the island were swept away.)

2. Research Methodology

There are various approaches to the TE in previous works, ranging from theorem proving to using linguistic resources such as WordNet [7]. Previously our approach is building a classifier with available textual features, such as the probability of alignment in monolingual machine translation of the input pair [8]. In this paper, we try to incorporate more observation into the SVM classifier [9]. The observation is made on the training set, and the feature selection is based on a ten-fold cross validation.

3. System architecture

The flowchart of our system is shown in Figure 1. The basic components are “preprocessing”, “word segmentation”, “Chinese character conversion”, ”special case filter”, “feature extraction”, ”sub-systems for special cases”, and “SVM” classifier.

3.1 Preprocessing

In this implementation, all the synonym or near synonym terms that in both t1 and t2 will be replaced to the same term, thus save the time of matching in later steps. This is a short cut to integrate semantic information into the system.

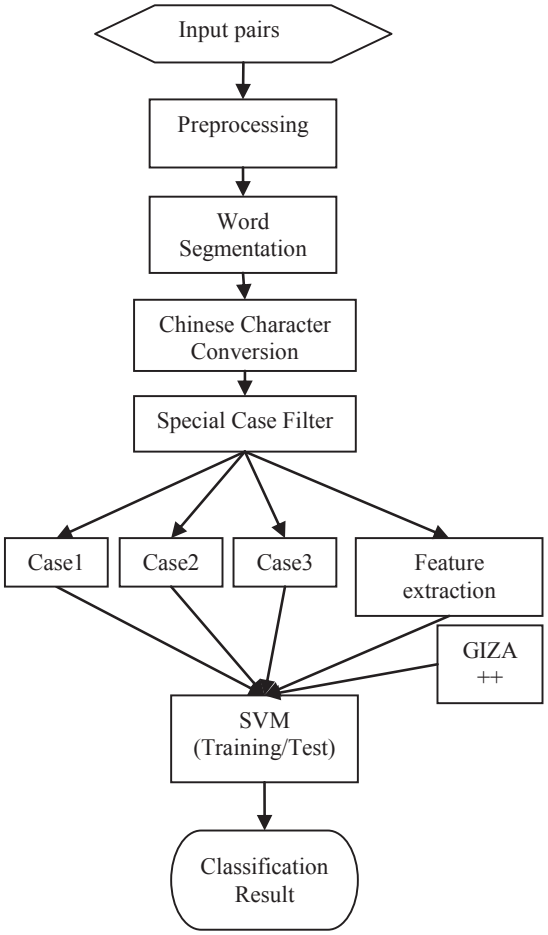

Figure 1. System flowchart

3.1.1 Normalizations

The normalizations in preprocessing include several small modifications on the terms that we regard them as the same term. For example, ” 葉望輝(Stephen J. Yates)” represent the name of a person in both Chinese and English, our system will normalize them into one common representation. Also, there are many different formats to represent time in Chinese such as shown in Table 2, our system will normalize them into the same representation. After the normalization, sentence with the same meaning but with different terms will be aligned easier. Thus, it can help our system to find features with higher accuracy.

Table 2. Examples of time expressions [15]

Type	Time expressions in text
Chinese only	一九九七年二月廿三日
Full type digit with Chinese	１９９７年２月２３日
Half type digit with Chinese	1997年2月23日
Digit only	1999-05-07
Duration	1999年延長至2001年

3.1.2 Background knowledge matching and substitution

The second part of our preprocessing system is to normalize synonym terms. The necessary knowledge can be collect from Wikipedia, HowNet[10], or TongYiCi CiLin [11].

3.2 Word Segmentation and Chinese character conversion

Since our system uses Stanford parser [12] to generate syntactical features, and based on the word segmentation result of ICTCLAS [13]. All the input sentences in traditional Chinese were converted into simplified Chinese by the online web service of Google translate [14].

3.3 Feature extraction

Table 3 lists the features used in our system, which are commonly used for the RTE task in previous works [15].

Table 3. Features Used in Our System

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)
11	Alignment score by GIZA++

The first three features are the number of common terms in both t1 and t2. The next three features are the BLEU scores [16][17]. The following for features are the numbers and differences of sentence lengths of t1 and t2. The last feature is the alignment score by GIZA++, which is the probability of how t1 can be a monolingual translation of t2.

3.3.1 Monolingual machine translation

Some recent research finds that machine translation can be a feature that helps RTE [18]. Our system also adopted such monolingual machine translation as a feature [6]. We believe that sentence with forward entailment relation must have the same meaning, so the second sentence can be translated from the first one. In our system, we use GIZA++ [19] as our monolingual machine translation tool. GIZA++ works according to the IBM model. In our system we use only the alignment score between the two input sentences, the formula is as follows:

$$p = \frac{\log\left\{\prod_{i,j=0}^{i,j=\max} p\left(t1_i \mid t2_j\right)\right\}}{n} \qquad (1)$$

3.4 Special cases in RITE-2 Chinese dataset

By observing the training corpus, we found that there are many special cases in the RITE-2 Chinese dataset. These cases are beyond the ability of our previous system. We decided to build a special case filter to pick them out and treated with special sub-systems.

Here we list first three cases:

Case1. The only difference is a negative word

We found that there are many pairs with almost identical words; the only difference is one sentence contains a negative word, such that the entailment relation of the two sentences becomes contradiction. In some cases, the only difference is an antonym.

Case 2. Inconsistent Temporal Information

Time is important information in many sentences, either reports a particular day or a particular duration. Therefore, two sentences with inconsistent temporal information either are independent or contradict each other.

Case 3.Inconsistent Numbers

Numbers in sentences often describe the quantity of objects. Just as the previous case, inconsistent numbers might imply either independence or contradiction.

Table 4 list examples of each case.

Table 4. Examples of special cases

Case	Example
Case 1	T1 若望保祿二世是四百五十多年來第一位**非**義大利籍的教宗
	T2 若望保祿二世是四百五十多年來第一位義大利籍的教宗
	The two sentences contradict to each other.
Case 2	T1 2003 年 4 月 **22** 日和平醫院爆發七位醫護人員集體感染 SARS, 4 月 24 日封院
	T2 2003 年 4 月 **24** 日和平醫院爆發七位醫護人員集體感染 SARS, 4 月 22 日封院
	The inconsistent temporal information causes a contradiction.
Case 3	T1 台、印雙邊貿易總額僅占台灣對外貿易總額 **12%**，顯示雙方經貿合作還有很大的成長空間。
	T2 台、印雙邊貿易總額僅占台灣對外貿易總額 **1.2%**，顯示雙方經貿合作還有很大的成長空間。
	The different number with the same context implies a contradiction.

Table 5 shows the numbers of pairs in the three special cases. The coverage is not very small in the data set, therefore, if our system can recognize the special cases and process them accordingly, the performance will increase.

Table 5. The number of pairs in the three special cases

	negative word	Inconsistent Temporal Information	Inconsistent Numbers
Training Set	15	43	42
Test Set	42	60	83

3.5 Support vector machine

The SVM tool used in our system is the LIBSVM [15], which can be used to train both binary class classifier and multiple class classifiers.

4. Experiment Result

In this section, we report the experiment results on training set and test set.

4.1 Formal run results

The formal run results of our system are shown in Table 6 and 7. We tested three different sets of features. In runs indexed 01, our system uses 10 features, which is the best configuration of our old system in RITE-1 [15]. For runs indexed 02, our system uses 11 features, the alignment score is treated as an additional feature [8]. For runs indexed 03, our system separately deal the special cases listed in section 3. Note that we use only the traditional Chinese training set in our training phrase, and use the same model to test both traditional Chinese test set and simplified Chinese test set.

Table 6. Formal run results of systems in RITE 2 CS task

Participants	BC	MC
bcNLP	73.84	56.82
MIG	68.09	44.74
CYUT	**67.86**	**42.52**
WHUTE	61.65	46.79
IASL	60.45	50.94
IMTKUTE	54.28	23.89
Yuntech	53.52	40.89
WUST	5014	40.87
JUNLP	48.49	24.38

Table 7. Formal run results of systems in RITE 2 CT task

Participants	BC	MC
IASL	67.14	46.32
MIG	67.07	45.15
IMTKUTE	65.99	32.36
WHUTE	65.55	45.50
Yuntech	62.31	40.14
KC99	57.67	42.16
CYUT	**53.52**	**26.26**
JUNLP	48.72	24.21
NTOUA	32.63	-
MCUIM	-	32.51

4.1.1 Formal run error analysis

The difference of official results and our system's best results are shown in Table 8 and 9. In the multi-class sub-task, our system tent to miss the contradiction and independent TE. It suggests that we should improve the recalls of contradiction and independent detection. In the binary-class sub-task, the numbers of Yes and No are quite near. It suggests that we should improve the precision.

Table 8. Multi-class comparison

	F	B	C	I
Official standard	277	145	106	253
CYUT-CS-MC-02	375	157	55	194

Table 9 . Binary-class comparison

	Y	N
Official standard	422	359
CYUT-CS-BC-03	429	352

Table 10 shows some error cases in our system's result. Case 1 is an independent pair, but our system misclassified it as bidirectional pair. The syntax and words of them are very similar. However, the subjects and the duration information are different. The error can be eliminated with careful collection of terminology and numerical information preprocessing.

Case 2 is a contradiction case, which our system misclassified it as bidirectional pair. There are two different transliterations of AIDS, which need to be normalized. Also, the numerical information is different. Case 3 is a common case that our system will misclassified it as forward pair just because all the words appeared in t2 are also appeared in t1. This kind of error require parsing technology to analysis the relations of terms in the sentence.

Case 4 indicate the requirement of more background knowledge. Since our system fail to infer that London means English in the pair, our system misclassified it as an independent pair. Case 5 shows that the antonym in a broader sense is necessary to recognize such case as contradiction rather than independent.

Table 10. Examples of error cases

Errors	Example pairs
Caes1	T1:**流感**病毒可在人体外存活**三到六**小时。
	T2:**冠状**病毒通常可在人体外存活**二到三**小时。
Caes2	T1: 大陆已有**四百万**人感染爱滋病。
	T2: 大陆有**八十五万**人感染艾滋病。
Caes3	T1: 美国奉行一中政策和遵守三公报的立场并未改变；切尼则进一步表示不支持台湾独立。
	T2: 美国不支持台湾走向独立。
Caes4	T1: 申奥成功的伦敦当局，在爆炸案后立即宣布取消庆祝活动。
	T2: 英国已停止所有庆祝申奥成功的活动。
Caes5	T1:我国生物技术可以与美国等先进国家相提并论，**解决**"异种核转殖"的问题并不难。
	T2: 我国生物技术可能**造成**异种核转殖等问题。

4.2 RITE4QA results

The best result of our system in the RITE4QA is shown in Table 11. The three settings are the same as in BC runs. Note that we use both the traditional Chinese and simplified Chinese training set in our training phrase, and use the same model to test both traditional Chinese test set and simplified Chinese test set.

Table 11. Best result of RITE4QA in RITE 2

Subtask	Top1
CYUT-CS-RITE4QA-02	7.33
CYUT-CT-RITE4QA-03	12.67

4.3 ADDITIONALRUNS

Although the dataset is almost the same and our system is almost the same. The difference of the performances in CT and CS of our system is quite significant. We find that the character conversion is the major cause. Several terms are not converted well, and the errors propagate to the word segmentation module and parser. A better CT to CS conversion tool will help to narrow down the gap.

The character conversion system used in the formal run was the online Google translate. In the additional runs, we use the CYUT character conversion system instead. The performance improved greatly as shown in Table 12. Thus, it is possible use one system on both CT and CS with a good character conversion system.

Table 12. Additional run results with different conversion systems

Character conversion system	BC	MC
Google Translate	53.52	26.26
CYUT system [21]	59.54	35.75

In section 3.4, we mentioned in special cases, a special processing unit might be design to some that special case. We show our result in Table 13. Each type can be improved significantly.

Table13. The BC accuracy of special cases

System	negative word	Inconsistent Temporal Information	Inconsistent Numbers
Formal run	52.38%	58.33%	53.01%
Special unit	71.42%	70%	72.28%

5. CONCLUSIONS AND FUTURE WORKS

This paper reports our system in the NTCIR-10 RITE-2 CT-BC, CT-MC, CS-BC, CS-MC and RITE4QA sub-tasks. Compared to the results of other teams [6], our system performs well in the CS-BC sub-task, where our best accuracy is 67.86%. However, the CT-BC sub-task, our best accuracy is much lower. In our additional run, we find that a better character conversion tool can help to boost the performance.

Compare to our old system in RITE-1, our new system used monolingual machine translation as a feature, and the results show that it can help to improve the performances in all CS runs. However, the treatment to the special cases did not show much improvement. We believe that special cases still need a separate process to deal with, but with more careful analysis. The additional run results show that special unit designed for special cases really help. If our analysis cover more special cases, the system can recognize textual entailment better. The system architecture with special unit is shown in Figure 2.

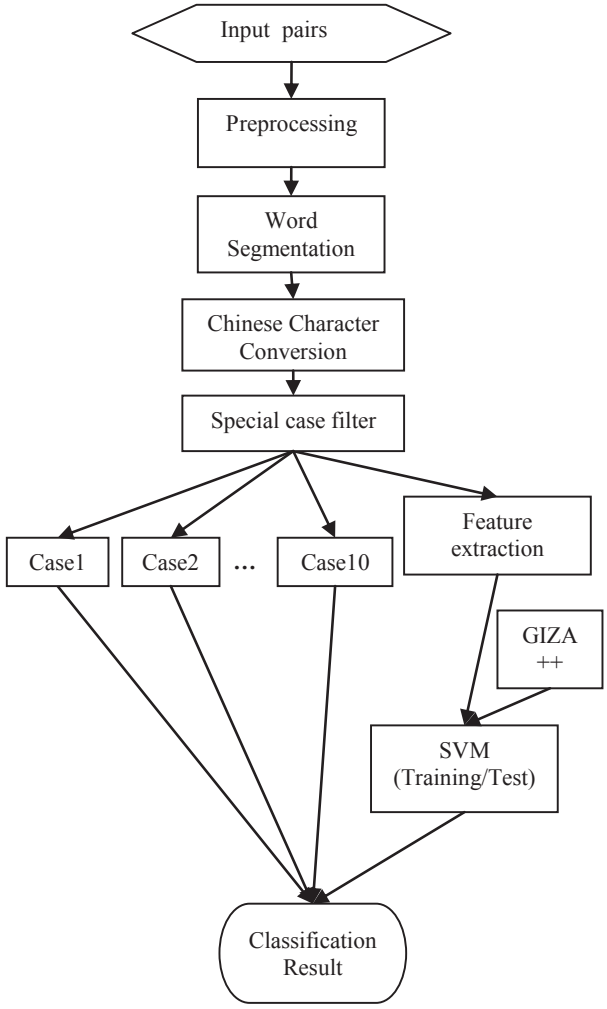

Figure 2. System flowchart with special unit

In the future, we will build more sub-systems on the special cases, as the architecture shown in Figure 2. Special cases might include:

1. synonym

We find that synonym and near synonym in each domain are crucial for RTE. Our preprocessing module do help the performance, more information are needed to reduce the cost of computation in the later stage.

2. Background knowledge on facts

There are many cases, even human can be wrong if that person does not have enough background knowledge. For example, in news, the capital of a state is used as a pronoun of the state. Without the knowledge, it is not possible to answer correctly. The knowledge might be collect from various encyclopedia.

3. Syntactical tricks

In several cases, the only difference between t1 and t2 is the syntax. The system should knowledge which word is the subject and which term is object. A parser is needed in these cases.

4. Polysemy and named entity

In some cases, the terms are the same in the pairs, but the meaning is different. Since any term can be used as a name in different context, it is also important where the term is indicating a named entity or not.

5. Negative modifier

Adding one negative modifier into a sentence or replace one term with its anatomy will make the sentence become a contradiction. However anatomy is hard to list in various domains.

6. Sentence reduction

In many forward pairs, t2 is a reduced sentence of t1. Carefully check the word and order might help to recognize such cases.

7. Logical inference

Entailment should be inference; several pairs do require logical inference. However, it requires a more rigorous normalization.

6. ACKNOWLEDGMENTS

This study is conducted under the "Digital Convergence Service Open Platform" of the Institute for Information Industry which is subsidized by the Ministry of Economy Affairs of the Republic of China. This research was partly supported by the National Science Council under NSC 100-2221-E-324 -025-MY2.

7. REFERENCES

[1] DAGAN, I., GLICKMAN, O., AND MAGNINI, B. 2006. The PASCAL recognizing textual entailment challenge.

[2] Ido Dagan and Oren Glickman, Probabilistic textual entailment: Generic applied modeling of language variability, In Proceedings of the Workshop on Learning Methods for Text Understanding and Mining, Grenoble, France,2004.

[3] Yongping Ou, Changqing Yao, "Recognize Textual Entailment by the Lexical and Semantic Matching", Computer Application and System Modeling, 2010 International Conference on V2-500 -504

[4] Dong-Bin Hua, Jun Dinga," Study on Similar Engineering Decision Problem Identification Based on Combination of Improved Edit-Distance and Skeletal Dependency Tree with POS", Systems Engineering Procedia Volume 1, 2011, Pages 406–413

[5] Ion Androutsopoulos and Prodromos Malakasiotis, "A survey of paraphrasing and textual entailment methods", Journal of Artificial Intelligence Research, Volume 38, pages 135-187, 2010.

[6] Yotaro Watanabe, Yusuke Miyao, Junta Mizuno, Tomohide Shibata, Hiroshi Kanayama, Cheng -Wel Lee, Chuan-Jie Lin , Shuming Shi, Teruko Mitamura, Noriko Kando, Hideki Shima, Kohichi Takeda," Overview of the Recognizing Inference in Text (RITE-2)at the NTCIR-10 Workshop", in Proceedings of the NTCIR-10 conference, Tokyo, Japan, 18-21 June., 2013.

[7] Christiane Fellbaum, "WordNet: An Electronic Lexical Database", The MIT Press, 1998

[8] Shan-Shun Yang, Shih-Hung Wu, Liang-Pu Chen, Wen-Tai Hsieh, and Seng-cho T. Chou, Improving Binary-class Chinese Textural Entailment by Monolingual Machine Translation Technology, in Proceedings of the IEEE IRI 2012, Las Vegas, USA,8 Aug, 2012.

[9] Prodromos Malakasiotis, Ion Androutsopoulos, "Learning textual entailment using SVMs and string similarity measures", In Proceedings of ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, pages 42–47, Prague, Czech Republic, 2007.

[10] Q. Liu, S.J. Li, Word Similarity Computing Based on How-net, Computational Linguistics and Chinese Language Processing，Vol.7, No.2, August 2002, pp.59-76

[11] Dong-Bin Hua, Jun Ding," Study on Similar Engineering Decision Problem Identification Based on Combination of Improved Edit-Distance and Skeletal Dependency Tree with POS",In Systems Engineering Procedia Volume 1, Pages 406–413, 2011.

[12] Stanford parser, http://nlp.stanford.edu/software/lex-parser.shtml

[13] ICTCLAS, http://ictclas.org/

[14] Google Translate, http://translate.google.com.tw/

[15] Shih-Hung Wu, Wan-Chi Huang, Liang-Pu Chen and Tsun Ku. Binary-class and Multi-class Chinese Textural Entailment System Description in NTCIR-9 RITE, in Proceedings of the NTCIR-9 workshop, Tokyo, Japan, 6-10 Dec., 2011.

[16] Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu, "BLEU: a method for automatic evaluation of machine translation", In Proceedings of the 40th Annual Meeting on ACL, pages 311–318, Philadelphia, PA, 2002.

[17] Liang Zhou, Chin-Yew Lin and Eduard Hovy, "Re-evaluating machine translation results with paraphrase support", In Proceedings of the Conference on EMNLP, pages 77–84, Sydney, Australia, 2006.

[18] Minh Quang Nhat Pham; Le Minh Nguyen; Shimazu, A. "Using Machine Translation for Recognizing Textual Entailment in Vietnamese Language", 2012 IEEE RIVF International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future .

[19] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models." Computational Linguistics, vol. 29, no. 1, pp. 19–51, 2003.

[20] LIBSVM, http://www.csie.ntu.edu.tw/~cjlin/lib

[21] Min-Hsiang Li, Shih-Hung Wu, Yi-Ching Zeng, Ping-che Yang, and Tsun Ku, Chinese Characters Conversion System based on Lookup Table and Language Model, Computational Linguistics and Chinese Language Processing, Vol. 15, No. 1, March 2010, pp. 19-36.