ISSN 1870-4069

Machine Translation Evaluation Metrics for Recognizing Textual Entailment

Tanik Saikh[1], Asif Ekbal[1], Debajyoty Banik[2], Pushpak Bhattacharyya[3]

[1]Indian Institute of Technology, Patna, India

[2]Kalinga Institute of Industrial Technology, Bhubaneswar, India

[3]Indian Institute of Technology Bombay India

debajyoty.banik@gmail.com, pb@cse.iitb.ac.in
{tanik.srf17, asif}@iitp.ac.in

Abstract. In this paper we propose a method for Recognizing Textual Entailment (RTE) that makes use of different machine translation (MT) evaluation metrics, namely *BLEU, METEOR, TER, WER, LE-BLEU, NIST* and *RIBES* and different versions of summary evaluation metric ROUGE, namely *ROUGE-N, ROUGE-S, ROUGE-W, ROUGE-L* and *ROUGE-SU* in a machine learning framework. Our main motivation of this paper is to investigate how MT evaluation metrics (which is generally used to judge the quality of an MT output), summary evaluation metrics (which is generally used to measure the quality of system generated summary) can be effective for determining TE relation between a pair of text snippets. Experiments on the datasets released as part of the shared task for recognizing textual entailment, RTE-1, RTE-2, RTE-3, RTE-4 and RTE-5 show the encouraging performance. We also performed a deeper comparative analysis of relevance of MT and summary evaluation metrics for the task of Textual Entailment (TE).

Keywords: Evaluation metrics, textual binding, automatic translation.

1 Introduction

One of the utmost challenging problems in the filed of Natural Language Processing (NLP) is to deal with language variability, that means there can be multiple ways to express a simple matter. Over the years researchers have been investigating a common framework which will be able to capture such language variability. Textual Entailment (TE) is an effective way to capture such language variability. Textual entailment requires complex linguistic analysis. TE was first introduced by [9] in the first track of recognizing textual entailment organized by *National Institute of Standard and Technology (NIST)*.

In this track TE was first defined as follows: suppose there are two texts fragments expressed as *Text (T)* and *Hypothesis (H)*. It is said that: ***T** entails **H** if, typically, a human reading **T** would infer that of **H** is most likely to be true*. For example, the text *T = "Mahatma Gandhi's assassin happened"* entails the hypothesis *H = "Mahatma Gandhi was died"*; obviously, if there exists one's assassin, then this person is died. Similarly, *T = "Mary lives in Germany"* entails *H = "Mary lives in Europe"*. On the other hand, *T = "Mary lives in Europe"* does not entail *H = "Mary lives in India"*.

There were many international conferences and evaluation tracks have been organized such as at Pattern Analysis, Statistical Modeling and Computational Learning (PASCAL)[4], Text Analysis Conferences (TAC)[5] organized by the United States National Institute of Standards and Technology (NIST), Evaluation Exercises on Semantic Evaluation (SemEval)[6], National Institute of Informatics Test Collection for Information Retrieval System (NTCIR) [7] since from the year of 2005.

These conferences and workshops produced many research articles which cover many approaches, varying from Lexical [9] [23], Syntactic [34] and semantics [4]. There are many applications in the field of NLP where TE can be employed, e.g. Machine Translation (MT) [16], Question-Answering (QA) [12] and Summarization [27] and many more. In MT evaluation, the machine generated output should entailed with the reference one.

In Question-Answering (QA), the answer produced by a machine must entail with that particular question. In summarization, the machine produced summary should entail with the reference one. Building an MT evaluation metric with the help of TE is a vital task. The proposed study makes use of various MT evaluation metrics and automatic summary evaluation metrics as features in machine learning framework.

1.1 Motivation

The MT evaluation metrics and automatic summary evaluation metrics are meant to investigate the quality of translation and summarization. It essentially does that by measuring similarity between the two (machine produced outputs and references) comparing piece of text fragments. We use these metrics to determine TE relationship. The study in [31] made use of very well established similarity metrics like Cosine, Jaccard, Dice, Overlap etc. and two MT evaluation metrics, namely BLEU [24] and METEOR [20, 1] for TE using RTE-1, RTE-2 and RTE-3 datasets.

This shows that MT evaluation metrics performed at per the ordinary similarity metrics in predicting the TE relation. Another study [30] made use of the same kind of similarity metrics along with the MT evaluation metrics on Indian languages (namely Hindi, Punjabi, Telugu and Malayalam) datasets to detect the paraphrase relation between a pair of sentences. The evaluation was performed on the datasets of the shared task *Detecting Paraphrases in Indian Languages (DPIL)* of *Forum for Information Retrieval Evaluation (FIRE-2016)*.

[4] http://pascallin.ecs.soton.ac.uk/Challenges/
[5] http://www.nist.gov/tac/tracks/index.html
[6] http://semeval2.fbk.eu/semeval2.php
[7] http://research.nii.ac.jp/ntcir/ntcir-9/

The study of [29] proposed a model which would be able to predict the TE relation between the two sentences (in RTE-1, RTE-2, RTE-3, RTE-4 and RTE-5 datasets) based on the MT evaluation metrics (BLEU, METEOR and TER) and one summary evaluation metric (ROUGE). Hence if BLEU, TER, and METEOR can take part in predicting TE relation between a pair of text snippets, there are other metrics too which could take part in deciding entailment relation between two piece of texts. The set of features applied here is very new to predict the TE relation in machine learning framework. There are also some works on *Microsoft Research Paraphrase (MSRP)* Corpus by [22] to detect paraphrases using MT evaluation metrics. So people are working in this line.

1.2 Related Work

There are umpteen number of research works carried out that could be found in literature on the datasets released in the shared task for recognizing textual entailment i.e. RTE-1, RTE-2, RTR-3, RTE-4 and RTE-5. Literature shows, [26] obtained the best accuracy in RTE-1 by the methods of word overlapping. They made use of BLEU, whose scores were used to assign "yes" or "no" class entailment decision based on some thresholds. The thresholds were learned based on some heuristics which were devised using the datasets. They obtained an accuracy of 70% on the dataset. The method defined in [15] achieved the best accuracy of 75% on RTE-2 dataset. The study proposed by the system made use of lexical and syntactic matching.

In RTE-3 dataset the best result was obtained in [14] using lexical alignment, knowledge extraction method, discourse commitment etc. The system proposed by [17] obtained the highest accuracy of 68.5%. The key concept of the participant's system is to map each word of hypothesis with one or more words of the text. They extensively made use of knowledge bases, namely DIRT, WordNet, VerbOcean, Wikipedia and Acronyms database etc. The system defined in [18] achieved the highest accuracy of 68.33% in RTE-5. The approach defined here is same as the approach defined in [17]. Additionally, they processed the LingPipe output with GATE in order to identify some of the named entity categories (e.g. nationality, language, job) which are within the scope of LingPipe.

Apart from the above cited works on TE and paraphrase detection there are few more works found in literature. The task described in [11] made use of BLEU, NIST, TER and Position independent word error rate (PER) to build a classifier which will be able to predict paraphrase relation between a pair of texts and also the entailment relation. They made use of *MSRP Corpus* and *RTE-1* for detection of paraphrase and entailment respectively. The work of [22] made use of *Microsoft Paraphrase Detection Corpus (MSRP)* and *Plagiarism Detection Corpus (PAN)* to re-examine the idea that automatic metrics which are generally used for judging the quality of a translation can also perform for the task of paraphrase detection. They used BLEU, NIST, TER, TERp [33], METEOR, SEPIA [13], BADGER [25], MAXSIM [5] metrics.

2 Feature Analysis

Features play a pivotal role in any machine learning assisted experiment. Hence identifying right combination of features which yield the best accuracy is the vital task. The following subsections define the features employed in the proposed study.

2.1 MT Evaluation Metrics

MT evaluation metrics generally used to judge the closeness between the machine translated output and the gold standard reference one. The more the closeness between them, the better is the translation system output.

Over the years, MT community proposed various metrics, namely BLEU [24], METEOR [20, 1], NIST [10], TER [32], Word Error Rate (WER) [37], Position independent word Error Rate (PER) and General Text Matcher (GTM) etc.We have incorporated almost all the metrics available in this study. We describe each of them in the following points:

1. BLEU: Bilingual Evaluation Understudy(BLEU) [24] is a metric which is perhaps the most popular MT evaluation metric developed by IBM. It measure the similarity between MT output and reference sentences by computing the n-gram precision between those sentences. Mathematically, it can be expressed as follows:

$$BLEU = BP \cdot \exp\left(\sum_{N}^{n=1} w_n log_{p_n}\right). \tag{1}$$

where, w_n is positive weights, summing to one; p_n is modified n-gram precisions. BP is brevity penalty computed as:

$$BP = \begin{cases} 1 & \text{if } c > r, \\ e^{\left(1-\frac{r}{c}\right)} & \text{if } c \leq r. \end{cases} \tag{2}$$

where, candidate translation length is c; and length of the effective reference corpus is r. Generally BLEU measures the similarity between the two sentences by computing n-gram matching. It can be termed as precision oriented metric. Similarity can be a way to detect TE relation, as it produce similarity, we wield this as a feature in this study.

2. LE-BLEU: *LE-BLEU* is also an MT evaluation metric which is based on n-gram matching, considered suitable for highly compounding languages [36]. It is an extension of BLEU, which consider fuzzy matches between two n-grams. It is supposed to better correlate with human judgment. The main drawback of BLEU is in exact n-gram matching. It could be possible not to match exact n-gram but having similar meaning. Such case LE-BLEU perform well than BLEU. LE-BLEU calculation is a kind of fuzzy calculation, whereas the calculation of BLEU is crisp kind of technique. So, there is great chance to get BLEU score zero, (if it is considering 4-gram matching, but machine up to 3-gram) where LE-BLEU provides satisfactory score. Thus, LE-BLEU can be a smart feature to recognize textual entailment between a piece of texts.

3. RIBES: Rank-based Intuitive Bilingual Evaluation Score (RIBES) [19] is also very useful for judging the MT output. Its calculation is based on significantly penalized word order mistakes and rank correlation coefficients. It is very effective for evaluating the accuracy score between distant sentence pairs. So, the similarity score between two distant sentences can easily predict their entailment.

4. NIST: The name NIST [10] came from *US National Institute of Standard and Technology*, which is used to evaluate between different text pairs. BLEU considers each n-gram to be of equal weight whereas NIST considers only the informative n-grams. NIST also differs from BLUE in its calculation of the brevity penalty in so far as small variations in translation length do not impact the overall score as much. As it also measures the similarity between the pair of text snippets, we exert this as feature.

5. METEOR: METEOR (Metric for Evaluation of Translation with Explicit Ordering) [20, 1] is a metric which measures the similarity between two sentences by computing the maximum-cardinality matching between those sentences. This match is used to compute the coherent based penalty. This computation is done by assessing the extent to which the matched words between texts constitute well ordered coherent "chunks". It considers lexical matching and synonym matching from the WordNet. As it takes both the approaches to measure the similarity into account, it could be an interesting feature in our study.

6. TER: Translation Error Rate (TER) [32] is also a metric to evaluate the performance of an MT output which is introduced in "Global Autonomous Language Exploration (GALE) Program" MT task. The central concept behind this metric is that edits required to change a hypothesis translation into reference translation. It generally produces the error rate by measuring the edit operations required to transfer the MT output to reference translation. Hence we get the similarity by taking complement of the error rate as shown in the following equation 3:

$$TER = \frac{\text{Number of edits}}{\text{average number of reference words}}. \tag{3}$$

Correctness is important to detect textual entailment, but it is not sufficient for this task. Error between two texts also important for recognizing textual entailment. So, we have considered TER and WER to fulfill its requirement.

7. WER: Word error rate (WER) [37] is also another very popular MT metric, which is also used in speech recognition. The metric works on word level and it is based on Levenshtein distance. It's computation is based on the minimum substitutions, deletions and insertions that have to be performed to convert the generated text into the reference text. It can be computed by the following formula 4: where *S:# of substitutions, D:# of deletions, I:# of of insertions, N:# of words in the reference*:

$$WER = \frac{S + D + I}{N}. \tag{4}$$

2.2 Summary Evaluation Metrics

Summary evaluation metrics are generally used to judge the quality of machine generated summary of a document which is generated automatically following an algorithm. *Recall-Oriented Understudy for Gisting Evaluation (ROUGE)* [21] is one of the most popular and widely accepted metric to evaluate summary. It comes up with it's several versions, we utilize all those versions here. Important ingredient for scoring for this metric is overlapping units. Units refer to n-gram word sequence word pairs between the hypothesis and reference sentences.

1. ROUGE-N: It computes the similarity score by measuring n-gram matching. Specifically, it is n-gram recall between a set of reference and candidate document. The mathematical equation of this metric is as follows:

$$ROUGE - N = \frac{\sum_{S \in \{\text{reference documents}\}} \sum_{\text{gram} \in}^{S} \text{Count}_{\text{match}}(\text{gram})}{\sum_{S \in \{\text{reference documents}\}} \sum_{\text{gram} \in S} \text{count(gram)}}. \quad (5)$$

where, count match(gram) is the maximum number of n-grams co-occurring in a candidate and a set of reference documents. The intuition behind using this metric is that it corresponds to the recall version of BLEU. So, we take into account precision from BLEU and recall from this metric.

2. ROUGE-L: It is basically the Longest Common Subsequence (LCS) based statistics. LCS is used as approximate string matching algorithm. To compare similarity between the hypothesis and reference documents normalized pairwise LCS is used in [28]. The higher the common matching between two texts the more the chances of the two texts to be textually entailed.

3. ROUGE-W: It is weighted LCS that favors consecutive LCSs. Problem with basic LCS is that it is unable to differentiate LCSs of different spatial relations. As it calculates similarity by taking weighted LCS into account, it can be an effective feature to predict entailment.

4. ROUGE-S: ROUGE-S is based on skip-bigram co-occurrence. It calculates the similarity score between two piece of texts, by considering bi-gram matching irrelevant of word order. We use this as a feature in our model.

5. ROUGE-SU: ROUGE-SU is the combination of skip-gram and unigram. It computes similarity by taking both of these into consideration. It is a different version to measure the similarity between a pair of texts snippets. This is also used as a feature.

3 Experimental Setup and Results

In this section we present preprocessing module, description of the datasets, experimental procedure, results, discussions and comparisons with the state-of-the arts.

Table 1. Statistics of the Datasets.

	# of T-H pair	
	Development	**Test**
RTE-1	567	800
RTE-2	800	800
RTE-3	800	800
RTE-4	0	1000
RTE-5	600	600

3.1 Preprocessing Module

Data are full of noisy by default. This performs the cleaning operation of such noise from the T-H pair contained in the datasets. We also removed the white spaces (if any) from the datasets. The example below shown is a T-H pair in the development set (taken from RTE-1).

```
<pair id="78" value="FALSE" task="IR"> <t>Clinton&apos;s
new book is not big seller here.</t>
<h>Clinton&apos;s book is a big seller.</h>
</pair>
```

Here `&apos;s` are replaced by "'" in the sentences, and then further it was converted into it's expanded form i.e. `Clinton&apos;s` converted into Clinton's.

3.2 Dataset

We use the datasets released in the shared task for recognizing textual entailment i.e. RTE-1, RTE-2, RTE-3, RTE-4 and RTE-5. The datasets of RTE-1, RTE-2 and RTE-3 correspond to the binary-class classification problem, whereas the datasets of RTE-4 and RTE-5 denote the ternary class classification problem. In our work we consider both binary and ternary classification. In Table 1 we show the number of T-H pairs present in the datasets.

3.3 Experiments

We extract T-H pairs from the datasets of RTE-1, RTE-2, RTE-3, RTE-4 and RTE-5. we calculate the similarity between T and H by exploiting the set of features that we already discussed. The scores obtained from each metric for a particular T-H pair are considered as feature value which are subjected for classifier's training and/or testing. As base learning algorithms, we use Support Vector Machine (SVM) [35, 6], Multilayer Perceptron (MLP) [2, 8], Logistic Regression [7] and Random forest (RF) [3]. We use the classifiers as available in weka[8]. The models are used to predict a class for an unknown T-H pair. We report the evaluation figures on the test set. The system predicts a class to each instances (T-H pair) of the test set. For RTE-4 we perform 10-fold cross validation results as we don't have access to the test dataset.

[8] http://www.cs.waikato.ac.nz/ml/weka/

Table 2. Results on test set of different datasets.

	Our System Accuracies(%)				Best Results(%)(Participant)
	SVM	Logistics	MLP	RF	
RTE-1	55.5	53.87	52.75	56.37	70 [26]
RTE-2	60.12	59.5	57	59.12	75 [15]
RTE-3	61.87	62.25	60.5	60.37	80 [14]
RTE-4	54.4	54.5	51.4	52.3	68.5 [17]
RTE-5	50	34.83	52.2	46.33	68.33 [18]

Table 3. Comparison results with baseline system.

Datasets	Baseline [29](%)	Proposed approach(%)
RTE-1	54	55.5
RTE-2	55	60.12
RTE-3	60	62.25
RTE-4	52	54.4
RTE-5	51	52.2

3.4 Results, Discussions and Comparisons

We evaluate all the three models on all the three datasets. We calculate *True Positive (TP), False Positive (FP), False Negative (FN)* and *True Negative (TN)*. We also report the accuracy of the system. We depict the accuracies obtained by the proposed system and the state-of-the-art models on different datasets in Table 2. In RTE-1 the best accuracy of our system is 56.37% using Random Forest (RF) compared to the best result reported as 70% by [26]. For RTE-2 we get an accuracy of 60.12% with SVM, whereas the best accuracy reported is 75% by [15]. For RTE-3, we obtain the highest accuracy of 62.25% in Logistics Regression framework, however the best accuracy reported in this track is 80% by [14]. For RTE-4, we obtain an accuracy of 54.5% using Logistic Regression classifier compared to the best reported value of 68.5% by [17].

The system with MLP yields an accuracy of 52.2% in RTE-5, whereas an accuracy of 68.33% was reported as the best result by [18]. It is to be noted that however we obtain relatively less accuracies compared to the state-of-the-arts, novelty of our proposed techniques is in the use of different MT and summary evaluation metrics for classifier's training for TE. Table 2 shows the evaluation results of different classifiers in our proposed system. The last column shows the state-of-the-art results obtained by the different participating systems in the respective tracks. For RTE-1 it is observed that Random Forest produces the best result among all the classifiers. For RTE-2, SVM model attains the best accuracy.

In RTE-3 and RTE-4, logistic regression model yield the highest performance. In RTE-5 MLP produces the best result. It is to be noted that different classifiers produces the best result for the different datasets. Comparisons to the baseline [29] models are presented in table 3. The system reported in [29] made use of all the datasets as what our proposed system exploits. The proposed system makes use of the baseline features as well as the others extracted from NIST, LE-BLEU, WER, RIBES, ROUGE-N, ROUGE-S, ROUGE-L, ROUGE-W and ROUGE-SU.

Table 4. Comparison between MT and Summary Evaluation metrics.

		Accuracy (%)	
Dataset	**Classifier**	**MT**	**Summary**
RTE-1	RF	56.25	50.37
RTE-2	SVM	58.37	60.5
RTE-3	Logistic	62.12	59.5
RTE-4	Logistic	55.1	51.5
RTE-5	MLP	53	50

Table 5. Features ablation study.

	BLEU	LE-BLEU	RIBES	NIST	WER	TER	METEOR	ROUGE-N	ROUGE-L	ROUGE-W	ROUGE-S	ROUGE-SU
RTE-1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
RTE-2	X	✓	X	X	X	X	N	N	N	X	✓	✓
RTE-3	N	X	X	X	✓	✓	N	X	N	X	X	X
RTE-4	N	✓	X	✓	X	N	✓	N	N	✓	✓	N
RTE-5	X	X	X	X	X	X	✓	✓	N	✓	X	N

This shows that our proposed approach is more effective than our baseline model which infers that usages of MT and summary evaluation metrics for the proposed task are, indeed, effective. We also performed a deeper analysis by comparing the result obtained by taking MT metrics alone as feature and the result obtained by taking summary evaluation metrics alone as feature. We run the model by the best performing classifier in each dataset and obtained and obtain two sets of results for MT and summary evaluation metric each. The results are shown in the following table:

3.5 Features Sensitivity Analysis

Features are very precious in any machine learning assisted experiment. In order to embellish the contribution of each feature to our predicting classes, we perform Ablation Study of features where we switch off a feature and then evaluate the model with the rest set of features and compare with the accuracy obtained by whole set of features (including the particular feature). Table 5 represents datasets in columns and row heading represent the features.

In this Table the "✓" represents that the corresponding feature has a positive effect on the performance, whereas "N" indicates that the particular features seem to have no effect and "X" denotes, the features seem to have negative effect on the performance. The table shows that, for RTE-1, all the features seem to have positive contributions. In RTE-2 *LE-BLEU*, *ROUGE-S* and *ROUGE-SU* features seem to be the most contributing features. The features, namely *METEOR*, *ROUGE-N* and *ROUGE-L* do not contribute significantly.

On the other hand, the features, namely *RIBES*, *NIST*, *WER*, *TER* and *ROUGE-W* seem to have negative effect on the performance. In RTE-3 datasets, only *WER and TER* are found to contribute more, *BLEU, METEOR and ROUGE-L* are found to be neutral, and *LE-BLEU, RIBES, NIST, ROUGE-N, ROUGE-W and ROUGE-S and ROUGE-SU* are found to be the features with negative effect.

Table 6. Results with contributing features only.

Datasets	Accuracies(%)
RTE-1	55.5
RTE-2	58.8
RTE-3	57.62
RTE-4	55.3
RTE-5	46.16

Table 7. Syntactic Parsing of T and H.

Syntactic Parsing	
T: John loves Merry.	H: Merry loves
(ROOT	(ROOT
(S	(S
(NP (NNP John))	(NP (NNP Merry))
(VP (VBZ loves)	(VP (VBZ Loves)
(NP (NNP Merry)))	(NP (NNP John)))
(. .)))	(. .)))

For RTE-4, *LE-BLEU, NIST, METEOR, ROUGE-W and ROUGE-S* are found to be the most effective features. For RTE-5, *METEOR, ROUGE-N and ROUGE-W* features contribute most. It is to be noted that, *LE-BLEU, METEOR, ROUGE-W and Rouge-S* are the features which are found to be the contributing features in most of the datasets. We perform another set of experiments by training and/or testing the classifier by considering only the contributing features. Results of these models are reported in Table 6. Please note that we report only the results of the best performing classifier (for each dataset).

3.6 Error Analysis

We perform error analysis to understand the shortcomings of our proposed method. Our system makes use of MT and summary evaluation metrics in a machine learning framework. Most of these metrics are based on lexical matching that may not be sufficient to capture the textual similarity always. These are not able to capture the syntactic and semantic ambiguities present in the corpus. Lexical matching ache from a drawback, sometimes it produces very high score for non-textually entailed text pair.

For the following example, *T: John loves Merry* and *H: Merry loves John*, n-gram matching produces a very high score and consequently, the system will mark that pair as entailed. Unigram and bigram matchings between T and H produce 3/3 = 1 and 0/3 = 0 scores, respectively. According to unigram matching the sentence pair is textually entailed, however they are actully should not be. On the other hand, if we parse T and H using a Stanford parser [9], it will produce the parsing information of that particular T and H as shown in the table 7.

[9] http://nlp.stanford.edu:8080/parser/index.jsp

Here none of the left or right child matches, hence they (T and H) are considered to be not textually entailed. Hence, syntactic information plays a vital role in determining TE relations. The system needs to be updated in this front. Let us consider the another example taken from the RTE-3 development set.

```
<pair id="251" entailment="YES" task="IR" length="short">
<t>Estimates vary widely, but it is believed there are up to
100 million children toiling in homes, factories, shops, fields,
brothels and on the streets of rural and urban India.</t>
<h>Child labor is widely used in Asia.</h> </pair>
```

There is only one common token (i.e **widely**) between T and H in the above example, so if we consider for lexical matching between those pair, the system will produce a very low score which is not sufficient to tag them (T-H pair) as textually entailed. However this pair should be defined as textually entailed. This needs further investigation.

4 Conclusion and Future Work

In this paper, we have proposed a system for recognizing textual entailment between a pair of text expressions which exploits MT evaluation metrics *namely BLEU, NIST, RIBES, LE-BLEU, TER, WER and METEOR* and summary evaluation metrics (*namely ROUGE-N, ROUGE-S, ROUGE-L, ROUGE-W and ROUGE-SU*) as features in a supervised machine learning framework. We develop models based on SVM, Logistic Regression, MLP and Random Forest.

Experiments performed on different benchmark datasets of RTE-1, RTE-2, RTE-3, RTE-4 and RTE-5 tracks show that our proposed approach attain encouraging performance. We believe that the proposed system is novel as the current study makes use of the features which are new of it's kind. The proposed system has great potential of applications for evaluating the MT and summary outputs.

The proposed study is in the opposite direction, which try to correlate MT and summary evaluation metrics with TE. The experiments reveal that MT and summary evaluation metrics which are generally use to judged the quality of machine produced translation and summary respectively, have a strong correlation which can also effectively take part in taking entailment decision between a pair of text snippets. Future works are directed towards the following dimensions:

- Will incorporate more such metrics in the existing system to build a more robust TE system.
- Will incorporate deep learning concepts in the existing system and want to make a comparative study.
- Planning to incorporate these MT and summary evaluation metrics in semantic textual similarity, which is another interesting problem to study.
- Planning to the work in reverse direction i.e. to build an MT evaluation metric by exploiting a robust TE system which will be based on deep learning approach.

References

1. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005). pp. 65–72 (2005)
2. Becerra, R., Joya, G., García Bermúdez, R. V., Luis, V., Rodríguez, R., Pino, C.: Saccadic points classification using multilayer perceptron and random forest classifiers in EOG recordings of patients with ataxia SCA2. Advances in Computational Intelligence. Lecture Notes in Computer Science , vol. 7903, no. 3 (2013)
3. Breiman, L.: Random forests. Machine Learning , vol. 45, no. 1, pp. 5–32 (2001)
4. Burchardt, A., Reiter, N., Thater, S., Frank, A.: A semantic approach to textual entailment: System evaluation and task analysis. In: Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. pp. 10–15. Association for Computational Linguistics (2007)
5. Chan, Y., Ng, H.: Maxsim: A maximum similarity metric for machine translation evaluation. Proceedings of ACL-HLT. pp. 55–62 (2008)
6. Chang, C. C., Lin, C. J.: Libsvm: A library for support vector machines. Association for Computing Machinery. Intelligence, Systems, Technology and Management , vol. 2, no. 3, pp. 1–27 (2011)
7. Collins, M., Schapire, R. E., Singer, Y.: Logistic regression, adaboost and bregman distances. Machine Learning , vol. 48, no. 1–3, pp. 253–285 (2002)
8. Costa, W., Garcia Fonseca, L. M., Sehn Körting, T.: Classifying grasslands and cultivated pastures in the brazilian cerrado using support vector machines, multilayer perceptrons and autoencoders. In: Machine Learning and Data Mining in Pattern Recognition - 11th International Conference. pp. 187–198 (2015)
9. Dagan, I., Glickman, O., Magnini, B.: The pascal recognising textual entailment challenge. In: Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment. pp. 177–190. Springer-Verlag (2006)
10. Doddington, G.: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proceedings of the Second International Conference on Human Language Technology Research. pp. 138–145. Morgan Kaufmann Publishers Inc. (2002)
11. Finch, A., Sook Hwang, Y., Sumita, E.: Using machine translation evaluation techniques to determine sentence-level semantic equivalence. In: Proceedings of the Third International Workshop on Paraphrasing. pp. 17–24 (2005)
12. Green, B. F., Wolf, A. K., Chomsky, C., Laughery, K.: Baseball: An automatic question-answerer. In: Papers Presented at the May 9-11, 1961, Western Joint IRE-AIEE-ACM Computer Conference. pp. 219–224. Association for Computing Machinery (1961)
13. Habash, N., Elkholy, A.: SEPIA: Surface span extension to syntactic dependency precisionbased MT evaluation. Proceedings of the NIST metrics for machine translation workshop at the association for machine translation in the Americas conference (2008)
14. Hickl, A., Bensley, J.: A discourse commitment-based framework for recognizing textual entailment. In: Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. RTE '07, Stroudsburg, PA, USA, Association for Computational Linguistics (2007). pp. 171—-176 (2007)
15. Hickl, A., Bensley, J., Williams, J., Roberts, K., Rink, B., Shi, Y.: Recognizing textual entailment with lccs groundhog system. In: Proceeding of the Second PASCAL Challenges Workshop. (2005)

16. Hutchins, W. J., Somers, H. L.: An introduction to machine translation. London: Academic Press (1992)
17. Iftene, A.: UAIC Participation at RTE4. Text Analysis Conference Workshop. National Institute of Standards and Technology pp. 17–19 (2008)
18. Iftene, A., Moruz, M.: UAIC participation at RTE5. Text Analysis Conference Workshop. National Institute of Standards and Technology (2009)
19. Isozaki, H., Hirao, T., Duh, K., Sudoh, K., Tsukada, H.: Automatic evaluation of translation quality for distant language pairs. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. pp. 944–952. Association for Computational Linguistics (2010)
20. Lavie, A., Agarwal, A.: METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In: Proceedings of the Second Workshop on Statistical Machine Translation. pp. 228–231. Association for Computational Linguistics (2007)
21. Lin, C. Y.: ROUGE: A package for automatic evaluation of summaries. In: Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004). pp. 74–81 (2004)
22. Madnani, N., Tetreault, J., Chodorow, M.: Re-examining machine translation metrics for paraphrase identification. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 182–190 (2012)
23. Pakray, P., Bandyopadhyay, S., Gelbukh, A.: Lexical based two-way RTE system at RTE-5. In: System Report, TAC RTE Notebook. (2009)
24. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: A Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics (2002)
25. Parker., S.: A New Machine Translation Metric. Proceedings of the Workshop on Metrics for Machine Translation at AMTA. (2008)
26. Perez, D., Alfonseca, E.: Application of the bleu algorithm for recognising textual entailments. In: In Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment. (2005)
27. Radev, D. R., McKeown, K. R.: Generating natural language summaries from multiple on-line sources. Computational Linguistics , vol. 24, no. 3, pp. 469–500 (1998)
28. Saggion, H., Radev, D., Teufel, S., Lam, W.: Meta-Evaluation of Summarization in a Cross–Lingual Environment Using–Based Metrics. Proceedings of COLING-2002, Taipei, Taiwan (2002)
29. Saikh, T., Naskar, S., Ekbal, A., Bandyopadhyay, S.: Textual Entailment using Machine Translation Evaluation Metrics. In: Computational Linguistics and Intelligent Text Processing - 18th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing 2017 (2017)
30. Saikh, T., Naskar, S. K., Bandyopadhyay, S.: JU_NLP@DPIL-FIRE 2016: Paraphrase Detection in Indian Languages - A machine Learning Approach. In: Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation, Kolkata, India. pp. 275–278 (2016)
31. Saikh, T., Naskar, S. K., Giri, C., Bandyopadhyay, S.: Textual entailment using different similarity metrics. In: Computational Linguistics and Intelligent Text Processing - 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, Proceedings, Part I. pp. 491–501 (2015)
32. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A Study of Translation Edit Rate with Targeted Human Annotation. In: In Proceedings of Association for Machine Translation in the Americas. pp. 223–231 (2006)
33. Snover, M., Madnani, N., Dorr, B., Schwartz, R.: TER-plus: Paraphrase, semantic, and alignment enhancements to translation edit rate. Machine Translation. , vol. 23, pp. 117—-127 (2009)

34. Vanderwende L., D. W.: What Syntax Can Contribute in the Entailment Task. In: Quiñonero-Candela J., Dagan I., Magnini B., d'Alché-Buc F. (eds) Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment. , vol. 3944, pp. 205–216. Springer (2006)
35. Vapnik, V. N.: The Nature of Statistical Learning Theory. Springer-Verlag New York, Inc. (1995)
36. Virpioja, S., Grönroos, S.-A.: LeBLEU: N-gram-based Translation Evaluation Score for Morphologically Complex Languages. In: WMT@ EMNLP. pp. 411–416 (2015)
37. Wang, Y.-Y., Acero, A., Chelba, C.: Is word error rate a good indicator for spoken language understanding accuracy (2003)