Analyzing Quality Estimation Of English-Hindi Machine Translation System

Nivedita Bharti, Nisheeth Joshi, Iti Mathur

Abstract: Automatically estimating the translation quality is a challenging topic of research in the field of MT. This paper describes the approach used for sentence level quality estimation problem on English-Hindi language pair. The purpose of the translation quality estimation (QE) is to predict a quality for unseen translated text without considering the reference translation. To perform the proposed technique, this submission conceived the quality estimation problem as a supervised learning approach. Feature extraction is an important step for supervised ML based quality estimation, and therefore, in this paper, we experimented with a set of multiple features along with the different ensemble type of learning algorithms. From the experimental results on the test set, we have found that Extra Tree based QE models gain improvements over the other two ensemble regressors. Moreover, the analysis of the performance evaluation measures show that the quality of the translation generated by the MT engine1 was best among all the four different MT engines.

Index Terms: Bagging, Extra Tree, Features, Machine Translation, Quality, Quality Estimation, Random Forest, Regression.

1 INTRODUCTION

In recent days, due to the introduction of neural machine translation, the MT has reached at the level of performance where it can be allowed for their integration into the real-world translation applications. Although, the outputs generated by the MT systems are rarely perfect and therefore requires human corrections. This issue has raised the need to estimate the translation quality as several different tasks: first, to predict the post-editing time or effort involved in correcting the generated translations. Second, to compare the translation outputs generated from several different MT systems as a ranking task. Third, to predict whether the obtained translation is published as it is, or it requires editing. The problem of quality estimation (QE) is different from the translation evaluation, as it does not rely on the reference translation. The QE approach only uses the information about the source language text, generated translation text, and the processes involved in obtaining the translation using the machine learning techniques. In this work, we have developed multiple QE models using different ensemble type of regression algorithms with an extracted set multiple feature. The rest of our paper is organized as: in the next section, we have described the work done by various researchers in the field of translation quality estimation. Section 3 describes the proposed method used for the development of our presented QE models. Experimental setup done in order to estimate the translation quality is outlined in Section 4. The experimental results are further discussed in Section 5. The last section concluded our presented approach.

- *Nivedita Bharti is currently a full-time research scholar at Banasthali Vidyapith, India., PH-09119110742. E-mail: nivedita2bharti@gmail.com*
- *Nisheeth Joshi is an Associate Professor at Banasthali Vidyapith, India, E-mail: nisheeth.joshi@rediffmail.com*
- *Iti Mathur is an Associate Professor at Banasthali Vidyapith, India, E-mail: mathur_iti@rediffmail.com*

2 RELATED WORKS

Earlier works on quality estimation have proved the consistency of the automatic predicted score with the human evaluation [1,2]. In this context, Specia & Farzindar [3] focused on automatically predicting the quality score for selecting the n-best candidate translations of the MT system. Later, Specia et al. [4] developed a quality prediction model corresponding to each translation system. After that the scores predicted by these individual models are used to rank each alternative translations of the same source sentence. He et al. [5] developed a binary classification model for selecting the sentence between two translation outputs. The first translation output was generated by statistical machine translation (SMT) system, and the second by using a translation memory. Subsequently, Specia et al. [6] concentrated on measuring the post-editing effort as the translation quality estimation task. Sánchez & Martínez [7] built a classifier which uses only the source side language information. Finally, the developed classifier selected which MT system should be used for translating a given source sentence. Although, the proposed approach showed a non-significant improvement. Avramidis et al. [8] did MT system evaluation on exploiting the learned ranking by using the parsing features, and without using the reference translations. Bach et al. [9] introduced linguistic and syntactic information to better estimate the translation quality as the dependency preservation checking. Besides, Avramidis [10] presented an approach of automatic ranking of multiple translation outputs at sentence level by exploiting the adequacy information as the features. These features were: nouns, punctuation occurrences, sentences and subordinate clauses. Avramidis [11] developed one distinct model for each HTER component and finally used the four individual predictions for computing the final quality prediction score, although the approach did not produce positive results. Wisniewskiet al. [12] built a random forest classifier with 16 features for predicting the translation quality of a word. In the same year, Souza et al. [13] trained two classification models by exploiting bidirectional long short-term memory (LSTM) recurrent neural networks and conditional random field (CRF) to develop the complete QE system at word level. Later, Tezcan et al. [14] trained regression model using baseline features in combination with the word level predictions as features for developing the sentence level ranking model.

Shah et al. [15, 16] used several neural features like neural network language model (NNLM) and word embedding features and then combined these neural features with the features generated by the QuEst++ framework [17] to finally predict the translation quality. Kim et al. [18, 19] applied RNN based neural MT model [20] for estimating the sentence level translation quality. Chen et al. [21] extracted neural features as well as the cross-entropy features for training the SVR model to finally build the quality estimation model. Recently, Etchegoyhen et al. [22] employed minimalist approach for sentence-level translation quality estimation model development. The approach used by the authors named as minimalist because it required few resources and minimal deployment efforts. Duma & Menzel [23] used a linear combination of tree and sequence kernels by considering the pseudo-reference translations as the data sources for estimating the translation quality. These kernel functions were not only applied on the source language text and the translated output text, but also on the back-translation and pseudo-references. The authors used partial tree kernel (PTK) as the tree kernel and subsequence kernel [24] as the sequence kernel. Kim et al. [25] proposed a novel bilingual BERT based translation quality estimation model at sentence and word levels by using the multi-task learning.

3 PROPOSED METHOD

To address the sentence level translation QE problem, the approach is broadly divided into two phases: feature extraction and the quality prediction. In the first phase, it extracts several features by exploiting the source language text, translated output text and the external language resources that covers the adequacy, fluency and the translation quality of the texts. Second phase predicts the translation quality by using the supervised machine learning based regression models. The detailed architecture of the followed approach is shown below in Figure 1.

Figure 1: *Proposed QE model architecture.*

4 EXPERIMENTAL SETUP

4.1 Dataset and Translation Outputs

To conduct the experimental works, we have used Tourism and Health domain English-Hindi parallel texts available online at "Technology Development for Indian Languages" (TDIL) site. We have collected 5,225 parallel texts and then randomly divide the dataset into training, development and test sets. Detailed corpus statistics is given below in Table 1.

TABLE 1
STATISTICS OF QE DATASET FOR ENGLISH-HINDI LANGUAGE PAIR

Dataset	Training		Development		Test
	en	hi	en	hi	En
# Sentences	4020		603		602
# Words	79,358	86,388	12,459	13,812	11,368
Unique Words	10,902	12,414	4249	4864	3812

The source (English) texts of the parallel corpus are used for obtaining the target translations by different MT systems, while Hindi language texts of the parallel corpus are considered as the reference or correct translations. To obtain the translation outputs of the given source sentence, we have used the four different MT systems as described below:

- Engine1: a neural web-based Google online translator developed by Google Inc.
- Engine2: another neural web-based Bing online translator developed by Microsoft Corporation.
- Engine3: Moses phrase based statistical machine translator [26].
- Engine4: Moses syntax based statistical machine translator [27].

4.2 Linguistic Tools

To perform the linguistic analysis of both source and translated output sentences in order to obtain the linguistic features, we have used several linguistic tools: Language modeling tool and POS tagger. Particularly, we have employed SRILM toolkit [28] for training the language models (LMs) of both source and translated output sentences with Kneser-Ney [29] smoothing technique. The POS tags for the English language are obtained using the Stanford POS tagger [30]. While for Hindi language, we have used HMM POS tagger developed by Joshi et al. [31].

4.3 Quality Labels

The quality of the translated outputs corresponding to all the given source texts present in the corpus are annotated manually on a 5-point scale using the Heval metric score [32] to prepare the training examples. This quality metric uses 5-point scale by considering the eleven linguistic parameters suggested by Joshi et al. [32]. The interpretations about the manually annotated quality score using the 5-point scale used for annotating the quality of the translation outputs is given in Table 2.

TABLE 2
INTERPRETATION OF THE 5-POINT QUALITY MEASURING SCALE

Score	Description
1	The translation is not acceptable
2	The translation is partially acceptable
3	The translation is acceptable
4	The Translation is perfect
5	The translation is ideal

4.4 Preprocessing

Since the sentences are basically available in the form of the raw data that ordinarily requires a conversion into the machine understandable form of representation. Therefore, we used tokenization and cleaning of both source and target sentences.

The preprocessing steps are:

- Tokenization: It produces a list of tokens or chunks by dividing the full sentence into tokens separated by a white space or a punctuation mark.
- Stop word Removal: It is a set of commonly used words. The reason behind the removal of this common words is that if we remove the commonly used words, we can focus on the important words.

- Stemming: It is used to obtain the root of the word in the sentence in order to reduce the noise present in the sentence.

4.5 Features

In this section, we have described the features that we have extracted from the source language and translated output sentences for training the quality prediction models and finally to make predictions on the unseen test set. Particularly, for our sentence level translation quality problem, we list all 32 features in order to build our QE regressor models. These features can be further categorized into the following broad categories:

POS features: This type of features is used to measure the text grammaticality.

- POS of the source word
- POS of the target word
- Percentage ratio of the Nouns/Verbs/Adjective in the source text
- Percentage ratio of the Nouns/Verbs/Adjective in the target text
- Punctuation symbols ratio in the source text
- Punctuation symbols ratio in the target text

Surface features: These are simple features which accounts for the difficulty involved in the translation task. These features are mainly extracted from the source texts. It includes features like:

- Words count
- Sentence length
- Out-of-vocabulary words count or the count of words that are not aligned
- unknown words count
- source to target length ratio

MT system features: This type of features commonly relied on the internal workings of the MT system and described the processes involved in generating the translation output.

- Phrase table
- Reordering model costs
- Weighted word penalty costs
- N-bests count for each source text
- Machine Translation output back-translation
- Word posterior probability

Language Model features: It accounts for the grammaticality and fluency of the target sentence.

- Unigram, bigram and trigram probability of the target texts
- Unigram, bigram and trigram perplexity of the target texts

IBM1 model scores: It is called as a bag-of-word translation model defined by Brown et al. [33]. This model generates the summation of all possible alignment probabilities between the given source text words and the target text words. The conditional probability distribution scores in both the directions are computed as:

- Source-to-Target IBM1 score
- Target-to-Source IBM1 score

Rule-based language quality check: Automatic rule-based language quality checks on both source and target language sentences provide a wide range of quality indications about the grammar, style, and terminology combinedly computed as an overall final quality score.

- Comma/ parenthesis+space
- Total errors
- Uppercase text start

4.6 Learning Algorithms and Evaluation Measures

We developed our QE models using the regression methods. Specifically, we applied three different ensemble-based regression algorithms: Bagging [34], Extra Tree [35] and Random Forest [36]. These learning algorithms are implemented using Scikit learn library [37] of python. The three algorithms are used with default parameters, whereas the RF parameters are optimized using grid search technique. Finally, to evaluate the developed QE models on the unseen test set, we have used the Mean Average Error (MAE), Root Mean Square Error (RMSE) and Pearson's correlation score as the evaluation measures.

5 RESULTS

This section demonstrates the experimental results of our proposed QE models. Particularly, Table 3 shows the examples of alternative candidate translations generated by the introduced translation systems, and their reference translation, corresponding to a given source sentence. While, Table 4-6 shows the scores predicted by our developed QE models on different MT engines by using different regression algorithms built with extracted features.

TABLE 3

EXAMPLE OF TRANSLATION OUTPUTS COMPARED WITH REFERENCE TRANLATION ACROSS TEST SET

S. No.	Source Sentence	Google Translation	Bing Translation	Moses Phrase based Translation	Moses Syntax based Translation	Reference Translation
1	The Taj is undoubtedly one of the most spectacular buildings of the world.	ताज निस्संदेह दुनिया की सबसे शानदार इमारतों में से एक है।	ताज बेशक दुनिया की सबसे शानदार इमारतों में से एक है।	The Taj undoubtedly है एक की सबसे spectacular buildings विश् व की ।	यहाँ विलेज है undoubtedly एक के के अधिकांश देखते बनता भवनों के के world.	ताज निस्संदेह दुनिया की सबसे शानदार इमारतों में से एक है।
2	The government of India office has more information	भारत सरकार के कार्यालय	भारत सरकार के कार्यालय के पास	The सरकार के India कार्यालय और	यहाँ वाशिन्दों भारत के दान है उतनी जानकारी	भारतीय सरकार के कार्यालय को दूसरे

	on other destinations as well.	को अन्य स्थलों पर भी अधिक जानकारी है।	अन्य गंतव्यों के बारे में भी अधिक जानकारी है।	अधिक जानकारी है पर के रूप में अन्य डेस्टिनेशन भी है।	पर अन्य , जैसे well.	स्थानों की भी अधिक जानकारी है।
3	French, British, Russian, and other Royalty were invited for the inauguration which coincided with the re-planning of Cairo.	फ्रेंच, ब्रिटिश, रूसी और अन्य रॉयल्टी को उद्घाटन के लिए आमंत्रित किया गया था जो काहिरा के पुन: नियोजन के साथ मेल खाता था।	उद्घाटन के लिए फ्रांसीसी, ब्रिटिश, रूसी और अन्य रॉयल्टी आमंत्रित किए गए थे जो काहिरा की फिर से योजना बनाने के साथ हुआ था।	French , British , Russian , और अन्य Royalty invited थे , जो के लिए inauguration coincided के साथ re-planning Cairo की है ।	French, British, Russian, और अन्य Royalty थे invited के लिए के inauguration जो coincided से के re-planning के Cairo.	फ्रांसीसी, अंग्रेज, रूसी तथा अन्य राजसी वंश के लोग इस उद्घाटन के लिए आमंत्रित किए गए थे जो काहिरा के पुनर्नियोजन के समय से मेल खाता था।

TABLE 4
QUALITY PREDICTION RESULTS OBTAINED BY MT ENGINE1 ON DIFFERENT REGRESSION METHODS

Regression methods	MAE	RMSE	Pearson's correlation
Bagging	0.0958	0.1346	0.6781
Extra Tree	0.0863	0.1183	0.7285
Random Forest	0.1112	0.1506	0.6323

TABLE 5
QUALITY PREDICTION RESULTS OBTAINED BY MT ENGINE2 ON DIFFERENT REGRESSION METHODS

Regression methods	MAE	RMSE	Pearson's correlation
Bagging	0.1114	0.1582	0.6014
Extra Tree	0.1107	0.1406	0.6654
Random Forest	0.1182	0.1656	0.5906

TABLE 6
QUALITY PREDICTION RESULTS OBTAINED BY MT ENGINE3 ON DIFFERENT REGRESSION METHODS

Regression methods	MAE	RMSE	Pearson's correlation
Bagging	0.1453	0.1783	0.4018
Extra Tree	0.1361	0.1709	0.4317
Random Forest	0.1636	0.2047	0.2957

TABLE 7
QUALITY PREDICTION RESULTS OBTAINED BY MT ENGINE4 ON DIFFERENT REGRESSION METHODS

Regression methods	MAE	RMSE	Pearson's correlation
Bagging	0.1518	0.1933	0.3141
Extra Tree	0.1499	0.1830	0.3330
Random Forest	0.2090	0.2618	0.1626

6 CONCLUSIONS

In this submission we provide a simple and efficient strategy based on ML for predicting the quality of the translation generated by the MT system at sentence level. We have evaluated the impact of different regression methods built on multiple features in combination. Using the presented features and the ensemble regressors, we have concluded that ET regressor appears out to be most effective regressor for translation QE problem and gives better results on all four MT engines. Moreover, QE model prediction ranked MT engine1 at first position, followed by MT engine2, whereas MT engine4 is ranked at last. The best setup on MT engine1 achieved a MAE of 0.0863, RMSE of 0.1183 and Pearson correlation of 0.7285 built using ET regressor.

7 REFERENCES

[1] C.B. Quirk, "Training a sentence-level machine translation confidence metric," LREC, pp. 825-828, 2004.

[2] M. Gamon, A. Aue, and M. Smets, "Sentence-level mt evaluation without reference translations: Beyond language modeling," Proc. Of European Association for Machine Translation (EAMT), pp. 103-111, 2005.

[3] L. Specia, & A. Farzindar, "Estimating machine translation post-editing effort with HTER," Proc. of the Second Joint EM+/CNGL Workshop Bringing MT to the User: Research on Integrating MT in the Translation Industry (JEC 10), Denver, pp. 33-41, Nov. 2010.

[4] L. Specia, D. Raj, & M. Turchi, "Machine translation evaluation versus quality estimation," Machine translation, vol. 24, no. 1, pp. 39-50, Mar. 2010.

[5] Y. He, Y. Ma, J. van Genabith, & A. Way, "Bridging SMT and TM with translation recommendation," Proc. of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 622-630, Uppsala, Sweden, July 2010.

[6] L. Specia, N. Hajlaoui, C. Hallett, and W. Aziz, "Predicting machine translation adequacy," Machine Translation Summit XIII, pp. 513-520, 2011.

[7] F. Sánchez-Martinez, "Choosing the best machine translation system to translate a sentence by using only source-language information", Proc. of the 15th Annual Conference of the European Association for Machine Translation, Leuven, Belgium, pp. 97-104, May 2011.

[8] E. Avramidis, M. Popovic, D. Vilar, & A. Burchardt, "Evaluate with Confidence Estimation: Machine ranking of translation outputs using grammatical features," Proc. of the Sixth Workshop on Statistical Machine Translation, Edinburgh, Scotland, pp. 65-70, July 2011.

[9] N. Bach, F. Huang, and Y. Al-Onaizan, “Goodness: a method for measuring machine translation Confidence,” Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Stroudsburg, PA, USA, vol. 1, pp. 211–219.

[10] E. Avramidis, “Comparative quality estimation: Automatic sentence-level ranking of multiple machine translation outputs,” Proc. of 24th International Conference on Computational Linguistics (COLING), Mumbai, India, pp. 115–132, 2012.

[11] E. Avramidis, “Efforts on Machine Learning over Human-mediated Translation Edit Rate,” Proc. of the Ninth Workshop on Statistical Machine Translation, Baltimore, Maryland, USA, pp. 302–306.

[12] G. Wisniewski, N. Pécheux, A. Allauzen, & F. Yvon, ”Limsi submission for wmt’14 qe task,” Proc. of the ninth Workshop on Statistical Machine Translation, pp. 348-354, June 2014.

[13] J.G.C. de Souza, J. González-Rubio, C. Buck, M. Turchi, and M. Negri, “FBK-UPV-UEdin participation in the WMT14 Quality Estimation shared-task,” Proceedings of the ninth Workshop on Statistical Machine Translation, Baltimore, Maryland USA, June, pp. 322-328, 2014.

[14] Tezcan, V. Hoste, B. Desmet, & L. Macken, “UGENT-LT3 SCATE system for machine translation quality estimation,” Proc. of the Tenth Workshop on Statistical Machine Translation, pp. 353–360, Lisboa, Portugal, Sep. 2015.

[15] K. Shah, R.W.M. Ng, F. Bougares, and L. Specia, “Investigating continuous space language models for machine translation quality estimation,” Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, pp. 1073–1078, Sep. 2015.

[16] K. Shah, F. Bougares, L. Barrault, & L. Specia, “Shef-lium-nn: Sentence level quality estimation with neural network features,” Proc. of the First Conference on Machine Translation, Berlin, Germany, vol. 2, pp. 838-842, Aug. 2016.

[17] L. Specia, G. Paetzold, and C. Scarton, “Multi-level translation quality prediction with quest++,” Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Beijing, China, pp. 115–120, July, 2015.

[18] H. Kim, HY Jung, H. Kwon, JH Lee, and SH Na, “Predictor estimator: Neural quality estimation based on target word prediction for machine translation,” ACM Transactions on Asian and Low-resource Language Information Processing, vol. 17, no. 1, pp. 1-22, Sep. 2017a.

[19] H. Kim, JH Lee, and SH Na, “Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation,” Proceedings of the 2nd Conference on Machine Translation, Copenhagen, Denmark, vol. 2, pp. 562–568, Sep. 2017b.

[20] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” ICLR 2015, pp. 1-15, 2015.

[21] Z. Chen, Y. Tan, C. Zhang, Q. Xiang, L. Zhang, M. Li & W.A.N.G. Mingwen, “Improving machine translation quality estimation with neural network features,” Proc. of the Second Conference on Machine Translation, Copenhagen, Denmark, vol. 2, pp. 551-555, Sep. 2017.

[22] T. Etchegoyhen, E.M. Garcia, & A. Azpeitia, “Supervised and Unsupervised Minimalist Quality Estimators: Vicomtech’s Participation in the WMT 2018 Quality Estimation Task,” Proc. of the Third Conference on Machine Translation, Brussels, Belgium, pp. 782-787, Oct. 2018.

[23] Duma M. and Menzel W., “The Benefit of Pseudo-Reference Translations in Quality Estimation of MT Output,” Proc. of the Third Conference on Machine Translation (WMT), Brussels, Belgium, vol. 2, pp. 789–794, Oct. 2018.

[24] R.C. Bunescu, & R.J. Mooney, “Subsequence kernels for relation extraction, “Advances in neural information processing systems, pp. 171-178, 2006.

[25] H. Kim, JH Lim, HK Kim and SH Na, “QE BERT: Bilingual BERT using Multi-task Learning for Neural Quality Estimation,” Proc. of the Fourth Conference on Machine Translation (WMT), Florence, Italy, vol. 3, pages 87–91, Aug. 2019.

[26] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, …, C. Dyer, "Moses: Open Source Toolkit for Statistical Machine Translation." Proc. of the 45th annual meeting of the ACL on interactive poster and demonstration sessions, Prague, pp. 177-180, June, 2007.

[27] H. Hoang, & P. Koehn, “Improved translation with source syntax labels,” Proc. of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, Uppsala, Sweden, pp. 409-417, July 2010.

[28] Stolcke, "SRILM-an extensible language modeling toolkit," In Proc. of the 7th International Conference on Spoken Language Processing (ICSLP 02), vol. 2, pp. 901–904, Denver, Colorado, USA, Sep. 2002.

[29] R. Kneser and H. Ney, “Improved backing-off for m-gram language modeling,” Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. I, pp. 181–184, Detroit, Michigan, May, 1995.

[30] K. Toutanova, D. Klein, C. D. Manning, & Y. Singer, “Feature-rich part-of-speech tagging with a cyclic dependency network,” Proc. of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology, vol. 1, pp. 173-180, Edmonton, Canada, May, 2003.

[31] N. Joshi, H. Darbari, & I. Mathur, “HMM-based POS tagger for Hindi,” Proc. of 2013 International Conference on Artificial Intelligence, Soft Computing (AISC-2013), pp. 341–349, 2013.

[32] N. Joshi, I. Mathur, H. Darbari, & A. Kumar, “HEval: Yet another human evaluation metric,” International Journal on Natural Language Computing, vol. 2, no. 5, pp. 21-36, Nov. 2013.

[33] P.F. Brown, V.J.D. Pietra, S.A.D. Pietra, & R.L. Mercer, “The mathematics of statistical machine translation: Parameter estimation,” Computational linguistics, vol. 19, no. 2, pp. 263-311, 1993.

[34] L. Breiman, “Bagging predictors,” Machine learning, vol. 24, no. 2, pp. 123-140, Apr. 1996.

[35] P. Geurts, D. Ernst, & L. Wehenkel, “Extremely randomized trees,” Machine learning, vol. 63, no. 1, pp. 3-42, Mar. 2006.
[36] L. Breiman, “Random forests,” Machine learning, vol. 45, no. 1, pp. 5-32, Oct. 2001.
[37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, ... & J. Vanderplas, “Scikit-learn: Machine learning in Python,” Journal of machine learning research, pp. 2825-2830, Oct. 2011.