Advances in Computer, Signals and Systems (2022)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2022.060709
ISSN 2371-8838 Vol. 6 Num. 7

Machine Translation Quality Estimation Algorithm Based on Intelligent Fuzzy Decision Tree Algorithm

Ruichao Li

School of Translation Studies, Xi'an Fanyi University, Xi'an, Shaanxi, China
280663560@qq.com

Keywords: Machine Translation, Intelligent Fuzzy Decision Tree Algorithm, Translation Quality Estimation, Phrase Structure Syntax Tree

Abstract: With the development and application of machine translation(MT) technology, MT results appear in more scenarios, but the translation quality cannot be guaranteed, and users need to know the quality of MT results to decide whether to adopt them or not. MT quality estimation (QE) is a key task in the field of MT, which can score the quality of a translation based only on the source language sentences and the MT. Unlike the methods of automatic translation evaluation, translation QE does not require the use of reference translation, which can save a lot of manpower and resources and is suitable for large-scale MT quality assessment scenarios without reference translation. In this paper, a phrase-structured syntactic tree is constructed based on an intelligent fuzzy decision tree(DT) algorithm, and the sentence-level and word-level QE results of MT translation are analyzed based on this tree structure. By comparing the training prediction results of translation QE models, it is obtained that the method of fusing dependent words with source language(SL) sequences is more helpful to the translation QE process than the method of fusing phrase structure features with source language sequences.

1. Introduction

With the vigorous development of computer science and technology, machine translation technology, in which computers replace human beings to translate between different languages, has gradually become mainstream. Since the MT technology is not yet mature, the automatic evaluation of the obtained translations is helpful for people to select the most suitable machine translation translation from them.

Since its introduction, the research on MT QE has gained extensive attention and made great progress. For example, some scholars propose a QE pseudodata construction method without introducing an additional MT system; they estimate the error distribution in the dataset by counting the existing translation quality, select sentence pairs in the parallel corpus, and construct translation errors in the target utterances. This method can construct pseudodata with similar distributions to the existing error distribution in the dataset without pretraining the model on a large-scale parallel corpus or training an additional machine translation system. However, in this method, the errors are falsified and the words associated with the errors are not the real words generated by the machine translation system, but randomly selected in the word list, which causes the constructed data to still

lack authenticity and make it difficult to simulate real translation scenarios [1-2]. The utility of grammatical information in English and French has also been found, arguing that tree kernels are a convenient and effective way to encode grammatical knowledge, and 144 features from the syntactic and dependency trees of the SL were selected to be combined with the linear and central features of the QE task [3]. Some studies used numerical syntactic features such as the maximum tree depth in the syntactic tree, and branch nodes in the tree to enhance the experimental results, and to determine the complexity of the sentence to facilitate later editing [4]. In summary, in machine learning-based translation QE methods, the effectiveness of the model mostly depends on the goodness of the features, and linguistic knowledge has been widely used in traditional translation QE studies with good results.

This paper firstly introduces the concept of fuzzy DT algorithm and its application in syntax tree construction in MT, then analyzes the data processing process of language model and language similarity, then constructs the translation QE system and system evaluation index, and finally analyzes the translation quality results by comparing the translation QE index values.

2. Basic Overview

2.1 Intelligent Fuzzy Decision Tree (DT) Algorithm

The fuzzy DT is constructed based on category attribute data, which is divided into training and validation data, and the model nodes can be considered as a set of attributes. The DT includes a root node and a leaf node, and the path from the root to each leaf represents a classification rule based on a hierarchical grouping [5]. DT algorithms such as ID3 algorithm, C4.5 algorithm have good classification effect, and the introduction of fuzzy set theory is to increase the classification performance of DT algorithm.

In this paper, we use this algorithm to construct a translation phrase structure tree for MT, which can be regarded as implicitly associated with the predicate theoretical meta-structure, and in the translation QE task, to capture the inherent structure of the SL as well as to avoid training difficulties and long-distance dependency problems due to the long extraction sequence, we extract the phrase component corresponding to each word, i.e., the parent node corresponding to each leaf node as the phrase syntactic labels are extracted [6].

Figure 1 shows a simple syntactic tree of the phrase structure.

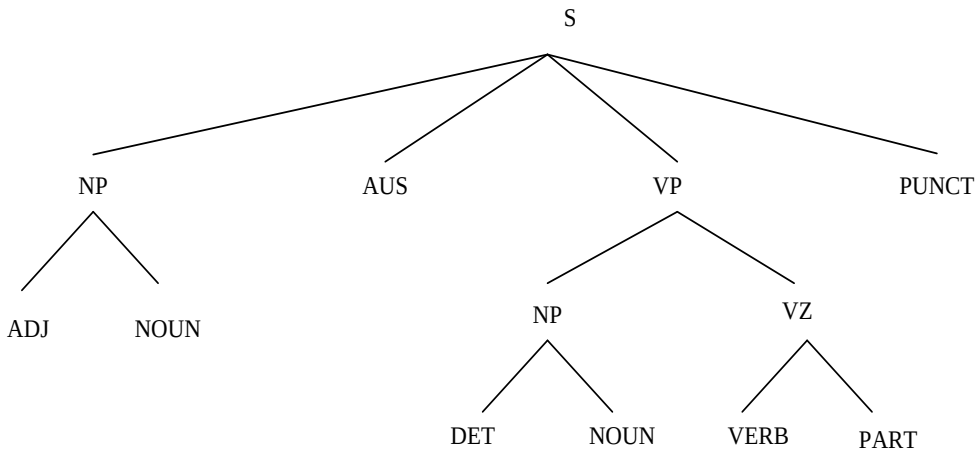

Figure 1: Phrase structure syntax tree.

2.2 Data Processing Based on Language Model and Utterance Similarity

(1) Data processing based on cut-off and generalization

An utterance in the actual translation application data is a natural language, and the original data cannot be directly used as the input of the MT model, so data conversion is required. The original application data is converted into information that can be used for result prediction through cut and generalization processing of the utterance, and such information is finally used as the input of the MT model to train the model or apply the model to translate the utterance [7].

(2) Data cleaning based on language models

No matter the duplicate data, noisy data or low-quality data will interfere with the training and learning of MT models, thus reducing the translation effect of the models, so effective data cleaning is an improvement of data quality, thus indirectly improving the MT effect. Therefore, data cleaning can be performed by evaluating utterances through constructing language models and analyzing utterance length ratios to ensure the quality of model input data to a certain extent [8].

(3) Data screening based on the similarity of utterances

The amount of data in the original application data is sufficient, but too large datasets for training models are time-consuming and resource-consuming, while one-sided sub-datasets do not characterize the whole well, so effective data screening is needed. By calculating the edit distance of utterance pairs to measure the utterance similarity and combining with probabilistic error-tolerant screening rules to complete data screening, a higher quality dataset for MT model training and learning is constructed [9].

2.3 Model Structure Applicable to the Translation QE Task

The translation QE task is a process of assessing the quality of MT translation using each word as the basic unit. Similar to the sentence-level task, the details of the word-level QE task are: the training phase provides the model with the input of the original text, the MT, the human postedited translation, and the tag file containing the OKBAD annotation for each word; in the testing phase, the model needs to give a 0K or BAD mark for the good or bad quality of each word with only the input of the original text and the corresponding MT [10]. The annotation of the tag file is also calculated by the TERCOM toolkit, which compares each word in the MT with the human postedited translation, and if the two words are consistent, it means that the corresponding MT is correctly translated and the translation is good, and is marked with a 0K tag; conversely, if the two words are not - consistent, it means that the MT has translation errors and the translation is poor, and is marked with A BAD label. Based on the different kinds of translation errors, the word-level translation QE task can be subdivided into translation word labeling (word level) and interword labeling (gap level), among which translation word labeling mainly reflects the translation errors of insertion and replacement, and interword labeling mainly reflects the translation errors of deletion [11-12].

3. MT Quality Assessment

3.1 Framework of Translation QE System

The main function of the system in this paper is to estimate the quality of translation produced by MT, so it is mainly divided into the following four parts, as shown in Figure 2.

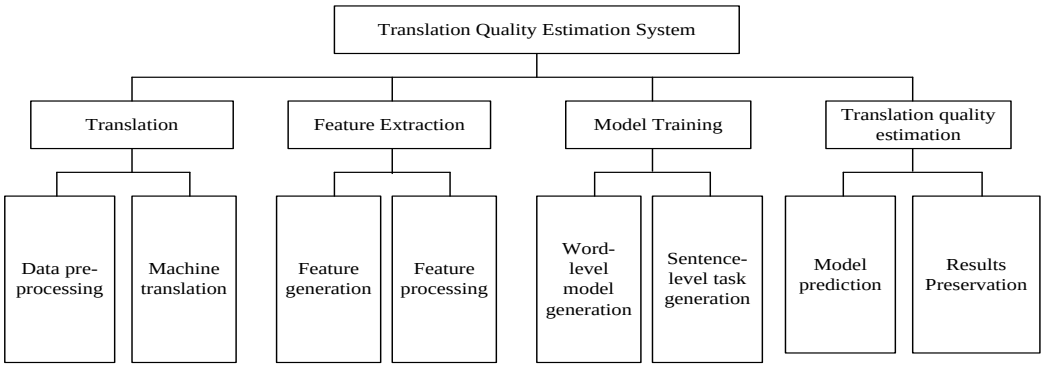

Figure 2: System structure.

As shown in the figure above, the four modules divide the work and work together to realize the completion of the translation QE process.

(1) Translation module

Data preprocessing module: Firstly, the source utterance is preprocessed, and for the Chinese and English utterances input by users, the words are firstly divided, and then the irregular behaviors in the sentence, such as full angle to half angle of punctuation, special symbols that do not conform to the utterance, etc. are adjusted, and then used as the input of the translation model. MT module: you can use your own trained MT model or call the existing translation API interface that has been made public, put the processed source language sentences into the translation model for translation, and generate the MT simply.

(2) Feature extraction module

Feature generation module: According to the multiple features mentioned above, the source language and the MT are processed differently. Feature processing module: The generated multiple features are processed separately for use in the subsequent translation QE model.

(3) Model training module

According to the extracted multiple features and feature fusion methods, experiments are conducted on different neural networks, mainly in two ways: single-feature experiment and multifeature combination experiment, and then the trained models are saved for subsequent result prediction.

(4) Translation QE module

Model prediction: The result prediction is to use the trained translation QE model to estimate the quality of the source language and machine-translated translations, and to make predictions according to different tasks. The sentence-level task uses the trained translation QE model to make predictions and obtains the HTER value corresponding to each MT (its physical meaning is the minimum editing distance between the MT and the correct translation after human modification, which takes a value between 0 and 1, and the closer to 0 means the better the translation is), and the word-level task also uses the neural network model to make predictions for each word in the MT and shows The word-level task also uses the neural network model to predict each word in the MT and display the corresponding word quality label in the translation. Result saving module: This module mainly saves the source language, MT and translation quality results into files separately.

3.2 System Evaluation Index

Word level is a classification task, that is, tagging each word in a sentence with a 0K/BAD tag. The evaluation metrics of the system can be evaluated by accuracy (P), recall (R), and F1 (F) values.

$$F = \frac{2 \times P \times R}{P + R} \times 100\% \quad (1)$$

By calculating the P, R, and F1 values of the 0K tag and the BAD tag, the product Fault of the F1 values of these two tags is then used as the word-level task evaluation index.

$$F_{mult} = F_{OK} \times F_{BAD} \quad (2)$$

FOK is the F1 value of the 0K tag and FBAD is the F1 value of the BAD tag. The word-level QE task aims to label correct target words as "OK"and incorrect target words as "BAD", so the word-level task can be regarded as a dichotomous problem.

4. Translation QE Results and Analysis

For the sentence-level(S-L) translation QE task, Pearson, Spearman, MAE and RMSE evaluation metrics are used, and the Pearson correlation coefficient and Spearman correlation coefficient are mainly evaluated, which mainly reflect the correlation between the machine predicted HTER value and the manual scoring. The results of the S-L experiments are given in Table 1.

Table 1: S-L QE results.

	Pearson's r	Spearman's ρ	MAE	RMSE
P-E(baseline)	0.542	0.514	0.125	0.163
P(+ PSG)-E	0.557	0.522	0.119	0.157
P(+ PSG)- E(+ PSG)	0.561	0.527	0.116	0.152
P(+ DSG)- E	0.573	0.531	0.114	0.150
P(+ DSG)- E(+ DSG)	0.579	0.538	0.111	0.148

Since there are relatively few "BAD" tags in the training corpus, the models can predict the result as "OK", so the experiment improves the overall performance (OP) of the system by increasing the F1 value of "BAD" tags. The OP of the system is improved by increasing the F1 value of "BAD" tags. The results of the word-level experiments are presented in Figure 3.

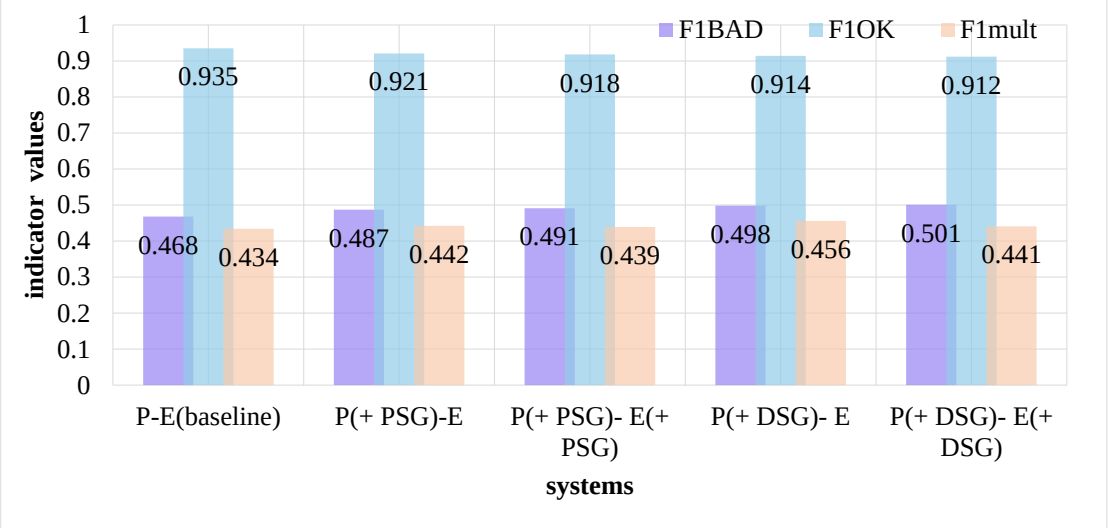

Figure 3: Word-level QE results.

In Table 1 and Figure 3, the system "P-E (baseline)" represents the baseline system (BS), which is built based on the original "predictor-estimator" framework with no other features added explicitly. System "P (a +PSG)-E" is based on the baseline system and encodes the syntactic

information (SI) of the SL phrase structure(PS) in the “predictor” module; the system “P(a +PSG)-E(+PSG)” is based on the “predictor” module, which is to add the SI of the SL PS in the “predictor” and “estimator” modules, respectively. Similarly, the system “P(a +DSG)-E” encodes the SL dependent SI in the “predictor” module; the system “P(a +DSG)-E(+DSG)” encodes the SL dependent SI in the “predictor” and “estimator” modules, respectively. The system “P(a +DSG)-E(+DSG)” is to add the source language-dependent SI in the “Predictor” and “Estimator” modules, respectively.

In Table 1 and Figure 3, the results of both the S-L and word-level(W-L) translation QE tasks are improved by adding SI. In the S-L task, the system that added syntactic knowledge to both the “predictor” and “estimator” modules performed best, with higher Pearson correlation coefficient scores compared to the BS. In the W-L task, the system that incorporated syntactic knowledge in the “predictor” module only performed the best and scored F1mult higher. The reason for this is that the PS syntax is more about the physical status of the components, while the dependency syntax reflects the deeper semantic modifications between the components, which can be more effective than the PS syntax. On the other hand, based on the strategy of fusing the syntactic feature sequences with the source language sequences, since the dependent words are extracted from the source language sequences and fused with them, the current words and their dependent syntactic features have already established the contextual association during the encoding process, so the fusion of their corresponding hidden states can make the two sequences more closely related.

5. Conclusions

MT is an important application direction in the field of natural language processing because it breaks through the communication barrier between different languages and is widely used and in great demand. And with the significant breakthrough in the research of deep learning, the field of MT has been rapidly developed, and the translation evaluation accompanying MT has also received attention. Therefore, through the syntactic decision tree established in this paper, the MT is evaluated by using the translation QE system, and the translation quality assessment results are outputted by the translation system, which is of great help for the subsequent study of translation QE.

Acknowledgements

This work was supported by the 2020 Annual Project of the 13th Five-Year-Plan for Education and Science in Shaanxi Province (Number: SGH20Y1509), and the research project of Xi’an Fanyi University (2021HZ-841)

References

[1] Binh Nguyen, Binh Le, Long H. B. Nguyen, Dien Dinh: PhraseAttn. (2022) Dynamic Slot Capsule Networks for Phrase Representation in Neural MT. J. Intell. Fuzzy Syst, 4, 3871-3878.

[2] Irene Rivera-Trigueros. (2022) MT Systems and Quality Assessment: A Systematic Review. Lang. Resour. Evaluation, 2, 593-619.

[3] Sainik Kumar Mahata, Avishek Garain, Dipankar Das, Sivaji Bandyopadhyay. (2022) Simplification of English and Bengali Sentences for Improving Quality of MT. Neural Process. Lett. 4, 3115-3139.

[4] Arda Tezcan, Veronique Hoste, Lieve Macken. (2020) Estimating Word-level Quality of Statistical MT Output Using Monolingual Information Alone. Nat. Lang. Eng, 1, 73-94.

[5] Jani Dugonik, Borko Boskovic, Janez Brest, Mirjam Sepesy Maucec. (2019) Improving Statistical MT Quality Using Differential Evolution. Informatica, 4, 629-645.

[6] Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, Sivaji Bandyopadhyay. (2022) Improved Neural MT for Low-resource English-Assamese Pair. J. Intell. Fuzzy Syst, 5,4727-4738.

[7] Ahmad Alos, Zouhair Dahrouj. (2020) Decision Tree Matrix Algorithm for Detecting Contextual Faults in Unmanned aerial Vehicles. J. Intell. Fuzzy Syst, 4,4929-4939.
[8] Subhashini Narayan, Jagadeesh Gobal. (2019) Optimal Decision Tree Fuzzy Rule-based Classifier for Heart Disease Prediction Using Improved Cuckoo Search Algorithm. Int. J. Bus. Intell. Data Min,4,408-429.
[9] K. Ramya, Yuvaraja Teekaraman, K. A. Ramesh Kumar. (2019) Fuzzy-Based Energy Management System With Decision Tree Algorithm for Power Security System. Int. J. Comput. Intell. Syst,2,1173-1178.
[10] Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindrich Helcl, Alexandra Birch.(2022) Survey of Low-Resource MT. Comput. Linguistics, 3, 673-732.
[11] Christian J. Park, Paul H. Yi, Hussain Al Yousif, Kenneth C. Wang. (2022) Machine vs. Radiologist-Based Translations of RadLex: Implications for Multi-language Report Interoperability. J. Digit. Imaging, 3, 660-665.
[12] Rajesh Kumar Chakrawarti, Jayshri Bansal, Pratosh Bansal.(2022) MT Model for Effective Translation of Hindi Poetries into English. J. Exp. Theor. Artif. Intell, 3,95-109.