Statistical Feature Aided Intelligent Deep Learning Machine Translation in Internet of Things

Yidian Zhang[3], Lin Zhang[3], Ping Lan[1], Wenyong Li[2,4], Dan Yang[2,5], and Zhiqiang Wu[1,2(✉)]

[1] College of Engineering, Tibet University, Lhasa 850000, China
lightnesstibet@163.com
[2] Center of Tibetan Studies (Everest Research Institute), Tibet University, Lhasa 850000, China
[3] School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, China
[4] School of Business Administration, Southwestern University of Finance and Economics, Chengdu 611130, China
[5] Beijing Foreign Studies University, Beijing 10089, China

Abstract. Internet of Things (IoT) networks have been widely deployed to achieve communication among machines and humans. Machine translation can enable human-machine interactions for IoT equipment. In this paper, we propose to combine the neural machine translation (NMT) and statistical machine translation (SMT) to improve translation precision. In our design, we propose a hybrid deep learning (DL) network that uses the statistical feature extracted from the words as the data set. Namely, we use the SMT model to score the generated words in each decoding step of the NMT model, instead of directly processing their outputs. These scores will be converted to the generation probability corresponding to words by classifiers and used for generating the output of the hybrid MT system. For the NMT, the DL network consists of the input layer, embedding layer, recurrent layer, hidden layer, and output layer. At the offline training stage, the NMT network is jointly trained with SMT models. Then at the online deployment stage, we load the fine-trained models and parameters to generate the outputs. Experimental results on French-to-English translation tasks show that the proposed scheme can take advantage of both NMT and SMT methods, thus higher translation precision could be achieved.

Keywords: Neural machine translation · Statistical machine translation · Neural network · Statistical feature extraction

1 Introduction

Internet of Things (IoT) networks enable the machines or devices to communicate with each other as well as humans. With IoT, humans could communicate

B. Li et al. (Eds.): IoTaaS 2020, LNICST 346, pp. 247–260, 2021.
https://doi.org/10.1007/978-3-030-67514-1_20

with machines, learn the machine's status, and achieve the intelligent control of IoT devices such as the fridge, the air conditioner, or the microwave ovens. These IoT services provide interactions between IoT devices and humans, while machine translation enables such interactions via providing the human-machine interaction interfaces for users.

Machine translation (MT) involves how to recognize the task of translation between two natural languages through a computer, which has been widely applied to the increasing social demands among people speaking different languages [11].

As a research branch of natural language processing (NLP), the MT systems confront the following difficulties: 1) The variability and the ambiguity of natural language: the same words, phrases, and sentences may represent different meanings in different contexts, while words, phrases or sentences with different or even opposite meanings can sometimes express the same meanings in the same context. Moreover, the mixed-use of new words and typos is also the major obstacle for the computer to understand natural language. 2) The difficulty of modeling: even if a unified linguistic rule can be provided to update and correct new words and typos promptly, it is challenging to build a mathematical model that can fully contain the above rules and meet the computer's affordability. 3) The corpus quality requirements: what kind of corpus can fully reflect the characteristics of language and how to collect such corpus are essential issues to be considered.

The neural machine translation (NMT) and statistical machine translation (SMT) are two main translation methods for achieving high-quality results. On the one hand, the deep learning network has been widely adopted in various research areas, such as the information retrieval [7], image processing [5] and speech recognition [2] etc. Based on the encoder-decoder architecture, the NMT method utilizes the deep learning network and models the translation process as "encoding & decoding".

The deep learning network could effectively adapt to the variability and ambiguity of natural languages and demonstrate outstanding performances to implement MT tasks. However, the NMT has to address the following issues: 1) Translation coverage [8]: a predetermined symbol (such as "EOS") is used as the end marker in the NMT process. The NMT model will finish the decoding process when the decoder generates this symbol, which cannot guarantee that all words in the source sentence can be translated, thereby inducing the problem of the "over-translation" or "under-translation". 2) Translation inaccuracy caused by the decoder [1]: attention mechanism is used in the NMT model, and the result of NMT is usually smoother than that of the SMT model thanks to the smoothing effect of attention weights. However, the smoothing processing may induce the loss of the sentence semantics. 3) Limited vocabulary [4]: the computation cost of the embedding layer in the encoder and the softmax layer in the decoder is directly proportional to the vocabulary. Therefore, the NMT model only uses limited words with higher generation probability and uses a

UNK symbol to represent other words. The occurrences of UNK symbols might result in the semantic truncation of both the input and output sentences.

On the other hand, SMT models the translation relationship between two languages by applying the statistical feature extraction technology, which could address the issues confronted by the NMT models. However, the SMT methods have the following defects: 1) Large model size: the SMT requires a lot of memory to store the statistical features extracted from the corpus, which leads to a larger model size than the NMT aided system. 2) Lack of fluency in results: due to the use of the invariant probability mapping, the SMT often produces accurate but not fluent results.

Against this background, in this paper, we propose to establish and realize a hybrid machine translation system by integrating the SMT model and the NMT model to improve the translation performances.

Different from existing methods, in our design, we propose to use the SMT model to score the generated words in each decoding step of the NMT model, instead of directly processing their outputs. These scores will be first converted into the generation probability corresponding to words by classifiers, and then be utilized to generate the output of the hybrid MT system.

Briefly, the major contributions of this paper include:

1. We propose to establish the unigram based word scoring system and the bigram based word scoring system based on the SMT model to unify the decoding granularity with the NMT model.
2. We propose establishing the communication channels between SMT and NMT models to unify the translation progress and realize the decoding guidance through these two word scoring systems.
3. We propose to construct classifiers and weighting units to integrate the decoding guidance and the generation probability into the NMT model.

The remaining part of this paper is organized as follows. Section 2 presents the details of our proposed system, including the framework of the proposed hybrid MT system model, the DL network structure and the word scoring system, and so on. Then Sect. 3 provides experimental results to validate our design. Finally, we conclude our findings in Sect. 4.

2 The Proposed Hybrid SMT-Aided NMT System

In this section, we will present the hybrid MT system model, and then introduce the word scoring systems and describe the classifier and the weight units.

Figure 1 illustrates the framework of our proposed hybrid MT system. In this system, we use the SMT model to provide the word generation guidance for the NMT model. Then we propose to use the unigram based word scoring system and the bigram based word scoring system to generate the word guidance. In the decoding process, the SMT model combines the attention weights shared by the NMT model to calculate word scores of each step and finally perform the decoding of the hybrid system with classifiers and the weighting unit.

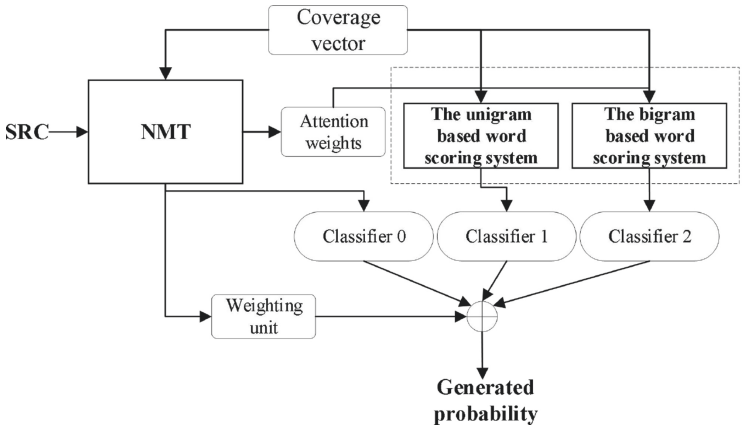


Fig. 1. The framework of the proposed hybrid MT system model.

2.1 Deep Neural Network for Hybrid MT System

In this hybrid SMT-aided NMT system, we proposed to apply the deep neural network (DNN) to achieve the MT. As illustrated in Fig. 2, the source sentence (SRC) is used as the input data set for both SMT and NMT models, while the output of this system is the sequence of word generation probabilities. The deep neural network consists of the input layer, the embedding layer, the recurrent layer, the hidden layer, and the output layer. Then for the DNN-based intelligent MT system, we utilize the word generation guidance of the SMT to combine the attention weights of NMT. Thus the reference output is the target sentence at the offline training stage and is a zero vector when the system is used for MT at the online deployment stage of the translation.

Moreover, to unify the NMT and SMT models' decoding progress, we introduce a coverage vector to mark the translated part of SRC explicitly. The length of the coverage vector is equal to that of the SRC, while elements in the vector are aligned with words in the SRC in order, which indicates that a word is untranslated when its corresponding element is 0 and is translated when the element is 1. To be more explicit, after each decoding step, if the generated word belongs to the word generation guidance, the corresponding element in the coverage vector will be updated to 1 following the setting given in [10].

2.2 The Unigram Based Word Scoring System

Figure 3 shows the unigram-based word scoring system, wherein the SRC is divided into discrete words and sent to the pre-scoring module. This module will calculate the pre-score of generated words according to the following equation:

$$SMT_1(y_t \mid y_{<t}, x) = \sum_{m=1}^{M} \lambda_m H_m(y_t, x_t) \tag{1}$$

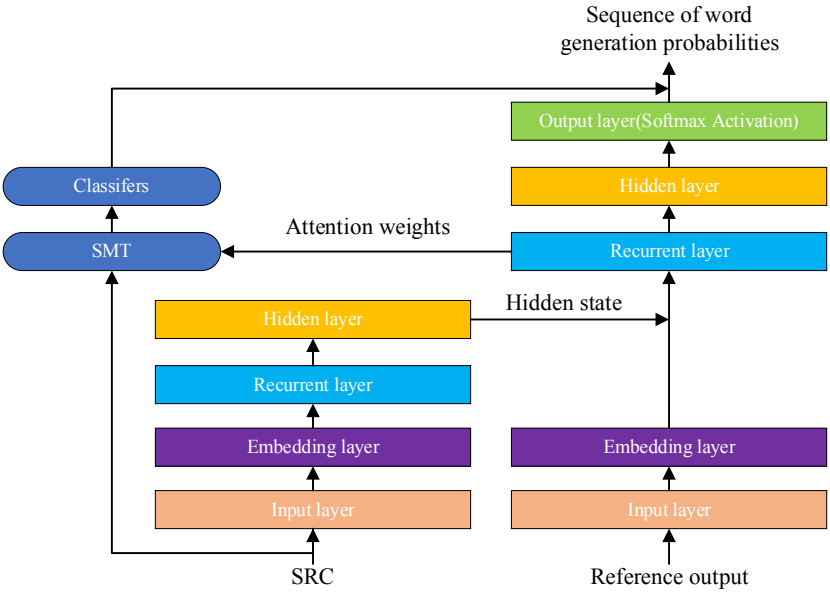


Fig. 2. The deep neural network of the proposed hybrid MT system model.

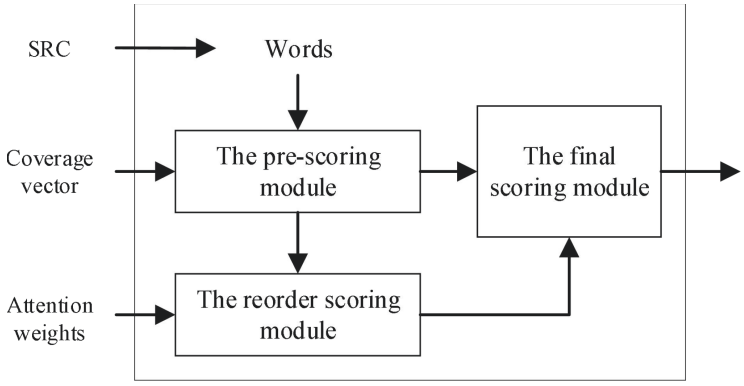


Fig. 3. The diagram of the unigram based word scoring system model

where x_t refers to the untranslated word in the SRC by the coverage vector, and y_t corresponds to the best n_{local} results to reduce the calculation burden of subsequent modules. H_m is a feature function used in the log-linear framework [6], and λ_m is the corresponding weight.

Considering that discrete words might result in the loss of word order, we propose to construct the reorder scoring module in the unigram based scoring system, which is used to compute the reorder score. With the same method presented by [9], the reorder score is calculated as follows:

$$d_1(y_t) = -\sum_{j=1}^{T_x} \alpha_{t-1,j} |sp_{y_t} - j - 1| \tag{2}$$

where $\alpha_{t-1,j}$ is the attention weights generated in the previous step of the NMT model, and sp_{y_t} denotes the position of source word which is aligned to the target word y.

Subsequently, we could establish the final scoring module for the unigram based word scoring system based on the pre-score and the reorder score of y_t. We apply a weighting unit in this module to adjust the pre-score and the reorder score weights, then the final score of y_t can be calculated as follows:

$$score_1(y_t) = \lambda \cdot SMT_1(y_t) + (1 - \lambda) \cdot d_1(y_t) \tag{3}$$

where $\lambda \in [0, 1]$ is the parameter of the weighting unit and needs to be used for the offline training. At last, the final score will be output to the classifier 1 for further processing.

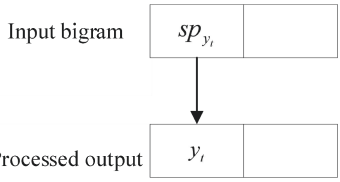
2.3 The Bigram Based Word Scoring System

To retain richer semantics of bigrams in SRC and improve the fluency of the SMT word generation guidance, we propose a bigram based word scoring system.

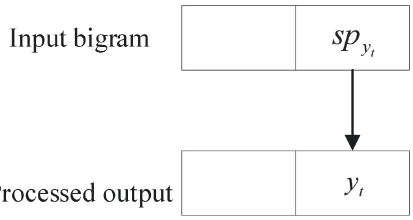
Unlike the unigram-based word scoring system, the source sentence x of length T_x is first divided into $(T_x - 1)$ bigrams in the bigram based word scoring system. Also, since the SMT model's output length corresponding to input bigram is not necessarily 2, the output will be complemented as a bigram when it is a unigram and keep only the first two words when its length is larger than 2.

To reduce the computational complexity, the output and the input bigram will adopt the alignment assumptions as given in Fig. 4. Namely, the first word in the output is generated at the first step and aligned with the first word of the input. The second word in the output is generated following the first word in the second step and aligned with the second word of the input.

To be more explicit, Fig. 5 presents the details of the alignment between the output sequence and the SRC, wherein y_t in the output sequence may be generated with two input bigrams. Next, we will present how to apply the above two assumptions to the bigram based word scoring system.



(a) y_t is the first word of the proposed output



(b) y_t is the first word of the proposed output

Fig. 4. The alignment assumptions for the input bigram and the output.

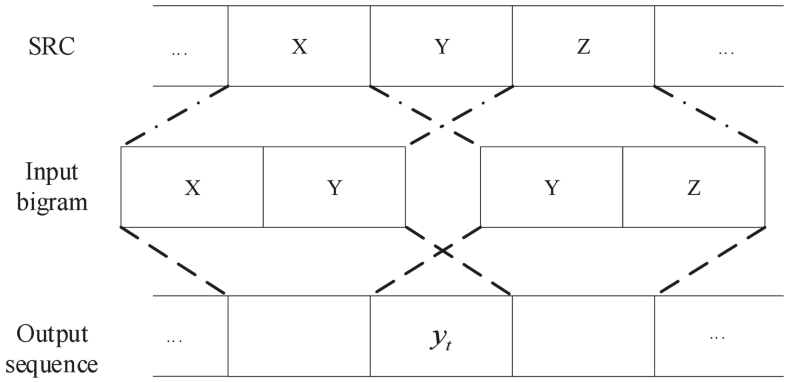


Fig. 5. The alignment between the output sequence and SRC

Figure 6 illustrates the framework of the bigram based word scoring system. Similar to the unigram based word scoring system, this system also consists of the pre-scoring module, the reorder scoring module, and the final scoring module, wherein the final score of the word is calculated based on the two assumptions.

For the first assumption, the pre-score is calculated by a similar method as that in the unigram based word scoring system:

$$SMT_{2-1}(y_t \mid y_{<t}, x) = \sum_{\langle y_t, \_\rangle \in \text{n-best}} \sum_{m=1}^{M} \lambda_m H_m(\langle y_t, \_\rangle, \langle x_t, \vec{x_t}\rangle) \tag{4}$$

where $\langle\cdot\rangle$ means the connections of the words in the brackets as the bigram, $\vec{x_t}$ is the last word of x_t in the SRC. $\langle y_t, \_\rangle$ indicates that the first word of the output

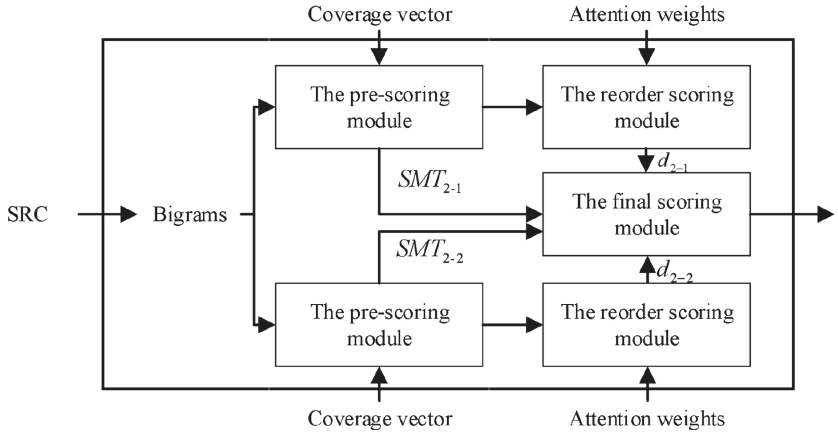


Fig. 6. The diagram of the bigram based word scoring system model

is y_t, while the second is limited by the coverage vector. Coverage vector limits the scoring range from both the input and the output, the limitations are as follows:

1) If the element in the coverage vector corresponding to x_t is 1, then $\langle x_t, \overrightarrow{x_t} \rangle$ does not meet input requirements.
2) If the element in the coverage vector corresponding to $\overrightarrow{x_t}$ is 1, then $\langle x_t, \overrightarrow{x_t} \rangle$ does not meet input requirements.
3) $\langle y_t, \_ \rangle$ only needs to meet the corresponding input $\langle x_t, \overrightarrow{x_t} \rangle$ limitations.

In this case, the reorder scoring is the same as Eq. 2, which is expressed as:

$$d_{2-1}(y_t) = -\sum_{j=1}^{T_x} \alpha_{t-1,j} |sp_{y_t} - j - 1| \tag{5}$$

For the second assumption, i.e., y_t is the second word of the output bigram, we calculate the pre-scoring based on the segment $y_{<t-1} = y_1, y_2, \ldots, y_{t-2}$ that has been generated in the previous step. Moreover, we revise the pre-scoring equation as:

$$SMT_{2-2}(y_t \mid y_{<t-1}, x) = \sum_{\langle \_, y_t \rangle \in \text{n-best}} \sum_{m=1}^{M} \lambda_m H_m(\langle \_, y_t \rangle, \langle \overleftarrow{x_t}, x_t \rangle) \tag{6}$$

where $\overleftarrow{x_t}$ is the previous word of x_t in the SRC, $\langle \_, y_t \rangle$ indicates that the second word of the output is y_t while the first one is limited by the coverage vector. It is worth mentioning that the limitations on scoring range are as follows:

1) If the element in coverage vector corresponding to x_t is 1, then $\langle \overleftarrow{x_t}, x_t \rangle$ does not meet input requirements.

2) If the current state of the element in coverage vector corresponding to $\overleftarrow{x_t}$ is not 1, then $\langle \overleftarrow{x_t}, x_t \rangle$ does not meet input requirements.
3) If the previous state of the element in coverage vector corresponding to $\overleftarrow{x_t}$ is not 0, then $\langle \overleftarrow{x_t}, x_t \rangle$ does not meet input requirements.
4) If the previous output of the hybrid system is $\bar{y}_{t-1}$, then $\langle \_, y_t \rangle$ which first word is not $\bar{y}_{t-1}$ does not meet the output requirements.

In this case, the reorder scoring module uses $\overleftarrow{x_t}$ as the basis for calculating the distance. Moreover, considering that the previous decoding step is occupied by the first word, the reorder score is calculated as follows:

$$d_{2-2}(y_t) = -\sum_{j=1}^{T_x} \alpha_{t-2,j} |sp_{y_t} - j - 2| \tag{7}$$

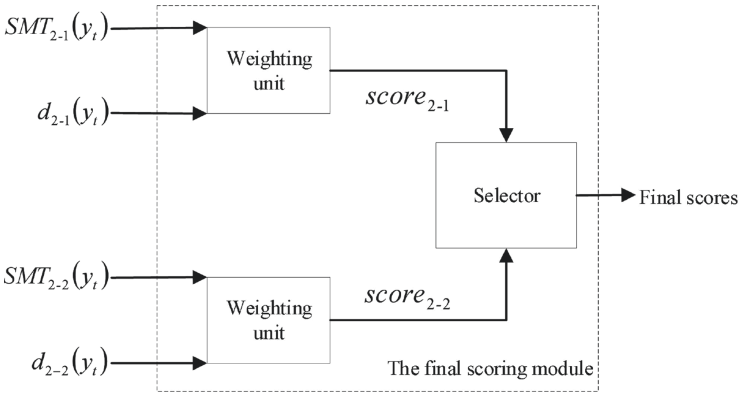


Fig. 7. The final scoring module.

Based on the above two sets of pre-scores and reorder scores, we propose a final scoring module for the bigram based word scoring system. As shown in Fig. 7, in the final scoring module. $score_{2-1}$ and $score_{2-2}$ are respectively calculated by weighting the two sets of scores as follows:

$$score_{2-1}(y_t) = \lambda \cdot SMT_{2-1}(y_t) + (1 - \lambda) \cdot d_{2-1}(y_t) \tag{8}$$

$$score_{2-2}(y_t) = \lambda \cdot SMT_{2-2}(y_t) + (1 - \lambda) \cdot d_{2-2}(y_t) \tag{9}$$

where λ is shared by the final scoring module in the unigram based word scoring system.

Finally, the selector will select the larger one of $score_{2-1}$ and $score_{2-2}$ as the final score of y_t. In particular, when y_t only appears in one assumption, the score under this case will be directly used as the final score of y_t.

2.4 Classifier and Weighting Unit

To unify the output structure of word scoring systems, we separate the Softmax layer from the NMT model and record it as the classifier 0. Moreover, we propose to denote the classifiers corresponding to the unigram based word scoring system and the bigram based word scoring system as classifiers 1 and 2. Furthermore, to match the output form of classifier 0, in both classifiers 1 and 2, we also adopt the Softmax function.

Moreover, to reduce the burden of the Softmax module, thereby decreasing the decoding complexity of the hybrid system, the classifier 1 first sorts all candidate words according to their final scores and only retains the best n_{global_1} results. Afterward, these scores will be mapped by the Softmax function, to achieve the conversion of the generated score to the generated probability. Additionally, the classifier 2 carries out the similar operations, while the only difference is the number of retained results, which is represented by n_{global_2}.

Besides, we propose to construct the weighting unit to control the weight of the output of each classifier, while each weight β_i corresponding to the classifier i is time-varying and calculated as follows:

$$\beta_i = \frac{\exp(g_i(s_t, y_{t-1}, c_t))}{\sum_{j=0}^{2} \exp(g_j(s_t, y_{t-1}, c_t))}, \quad i = \{0, 1, 2\} \tag{10}$$

where $g_i(\cdot)$ is the *sigmoid* function.

2.5 Offline Training and Online Deployment

At last, we briefly introduce the training and the online deployment of the SMT aided NMT system. The parameters shared between the proposed system and the NMT model are initialized by the fine-trained model. The other parameters of the hybrid system are randomly initialized.

After that, all parameters will be trained to minimize the negative log-likelihood of the variables:

$$L(\theta) = -\frac{1}{N_{train}} \sum_{n=1}^{N_{train}} \sum_{t=1}^{T_y} \log(p({y_t}^n \mid {y_{<t}}^n, x)) \tag{11}$$

where N_{train} is the number of bilingual sentence pairs in the training data and T_y is the length of expected output y.

Then at the online deployment stage, with the trained DNN model, the proposed intelligent MT system could output the translation results in real-time.

3 Numerical Results and Analysis

In this section, we provide numerical results to analyze the performances of the proposed intelligent MT system with the paired French-to-English data sets. In the analysis, the parameter settings are given as follows.

In the numerical analysis, we use WMT2013 as the development set; then, we test the proposed system on newstest2008, newstest2009, newstest2010, newstest2011 in WMT2012. We then use Moses and RNNSearch as the benchmark systems for translation performance comparisons for the proposed intelligent SMT aided NMT system. For Moses, the preprocessing of the bilingual corpus adopts the default method, and use KenLM [3] to train a 4-gram language model by exploiting the target corpus and then realize the grammatical control of the output. Besides, We run Giza++ for word-aligning ours parallel corpus, and carry out the RNNSearch with an open-source NMT system GroundHog, wherein the system setting follows that given in [9].

Additionally, in the unigram based word scoring system, n_{local} is set to 5. That is to say, for each input, the pre-scoring module keeps only the best 5 results. In the bigram based word scoring system, n in $n-best$ is set to 3, while n_{global_1} in classifier 1 is set to 20 and n_{global_2} in classifier 2 is set to 10.

3.1 Translation Performance Analysis

We first analyze the translation performances of the proposed system and compare them with the benchmark RNNSearch system. As shown in Table 1, we compared three examples generated by RNNSearch and the proposed system. In the first example, the fragment "Am érique du nord" at the end of SRC means "North America". From the translation results, it can be seen that RNNSearch

Table 1. Translation examples generated by the RNNSearch and the proposed systems

SRC	Hrafnsson réagissait manifestement aux déclarations faites en Amérique du nord
RNNSearch	Hrafnsson was clearly reacting to statements from North America in the north
The hybrid system	Hrafnsson was clearly reacting to statements from North America
SRC	Les hommes qui ont l'index plus long que l'annulaire sont exposés à un risque moindre d'avoir un cancer de la prostate
RNNSearch	Men who hand around the longer longer than ring finger face may lower off prostate cancer
The hybrid system	Men who have taken a longer index finger than ring finger are at lower risk of prostate cancer
SRC	Un homme et une femme d'une taille de 180 cm ont donc besoin d'un lit de 210/220 cm
RNNSearch	Man and woman with a height of about future gains need a bed of UNK cm
The hybrid system	Man and woman with a height of about 180 cm then need a bed cm

Table 2. Statistics of the percentages of UNK symbols for RNNSearch and the proposed systems

System	newstest2008	newstest2009	newstest2010	newstest2011	Average
RNNSearch	4.93%	5.18%	5.33%	5.38%	5.21%
The proposed system	3.81%	4.24%	4.52%	4.42%	4.25%

has the over-translation problem due to the excessive interpretation of the meaning of "north", while the proposed intelligent MT system demonstrates better control of the translation performances. In the second example, the RNNSearch based MT system still suffers from the over-translation and inaccuracy problems, while the proposed system achieves higher accuracy of translations. For the third example, we could notice that the proposed system is capable of expanding the vocabulary while retaining more SRC semantics. To be more explicit, in this example, "180 cm" is retained in the result of the hybrid system but not appearing in the output of the RNNSearch based system.

Furthermore, Table 2 compares the statistics of the UNK symbols in the proposed systems with that of the benchmark RNNSearch based system by using the test sets. It can be noticed from the table that both RNNSearch and our systems suffer from the limited vocabulary problem. It is worth pointing out that the UNK symbols in the proposed system are significantly fewer than those in RNNSearch system. This fact indicates that the proposed intelligent SMT aided NMT system could effectively mitigate the limited vocabulary effects on the performances of the intelligent DNN aided design.

Table 3. The BLEU results on French-to-English translation task

System	newstest2008	newstest2009	newstest2010	newstest2011	Average
Moses	21.04	23.60	24.88	20.48	22.50
RNNSearch	21.00	23.47	25.14	21.48	22.77
The proposed system	23.41	26.64	28.63	23.83	25.63

Moreover, Table 3 compares the BLEU results for completing the French-to-English translation task. It could be observed that the proposed system outperforms the benchmark RNNSearch system and the benchmark Moses system by 3.13 BLEU points and 2.86 BLEU points.

Table 4. Details of BLEU scores on newstest2011

System	1-gram	2-gram	3-gram	4-gram	BLEU
Moses	57.1	26.5	14.2	8.2	20.48
RNNSearch	50.4	25.0	16.4	12.2	21.48
The proposed system	52.9	28.1	19.7	15.6	23.83

Finally, Table 4 shows details of BLEU scores on newstest2011. It can be seen that both 1-gram and 2-gram scores of Moses are higher than that of RNNSearch, which are 6.7 and 1.5, respectively, reflecting the higher accuracy of Moses in unigrams and bigrams. Meanwhile, RNNSearch's 3-gram and 4-gram are higher, which are 4.2 and 4.0, respectively, reflecting its higher fluency in translation. The proposed system fully combines the advantages of accurate translation of Moses and smoother translation results of RNNSearch, except for 1-gram score is lower than Moses, other scores are higher than those of both Moses and RNNSearch, and the highest BLEU score is obtained finally.

4 Conclusions

In this paper, we propose an intelligent deep learning machine translation mechanism based on the extracted statistical feature to achieve the human-machine interactions for IoT devices. In the proposed MT system, the SMT subsystem combines the current state to calculate the word generation score for each decoding step and guides the translation process. Namely, we employ the SMT to score the generated words and then convert the scores to the generation probability. After the offline training, the resultant well-configured neural networks could output the translation results at the online deployment stage. The experimental results demonstrate that our proposed SMT aided intelligent MT system can effectively improve the translation performances. With the proposed design, IoT devices could achieve more effective interactions with humans, thereby greatly enhancing user-friendliness performances.

Acknowledgements. This work was supported in part by the State Key Program of National Social Science of China (No. 18AZD035), the Key Research & Development and Transformation Plan of Science and Technology Program for Tibet Autonomous Region (No. XZ201901-GB-16), the Special Fund from the Central Finance to Support the Development of Local Universities (No. ZFYJY201902001) and the National Natural Science Foundation of China (No. 71964030).

References

1. Arthur, P., Neubig, G., Nakamura, S.: Incorporating discrete translation lexicons into neural machine translation. In: Proceedings of Conference on Empirical Methods in Natural Language Processing (2016)
2. Dahl, G., Yu, D., Deng, l., Acero, A.: Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio Speech Lang. Process. **20**(1), 30–42 (2012)
3. Heafield, K.: KenLM: faster and smaller language model queries. In: Proceedings of the Sixth Workshop on Statistical Machine Translation, WMT 2011, pp. 187–197. Association for Computational Linguistics, USA (2011)
4. Jean, S., Cho, K., Memisevic, R., Bengio, Y.: On using very large target vocabulary for neural machine translation. In: Proceedings of 53rd Annual Meeting of the Association for Computational Linguistics. 7th International Joint Conference on Natural Language Processing, Beijing, China, vol. 1, pp. 1–10. Association for Computational Linguistics, July 2015

5. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. In: Proceedings of Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
6. Och, F., Nev, H.: Discriminative training and maximum entropy models for statistical machine translation, pp. 295–302. Association for Computational Linguistics, Philadelphia (2002)
7. Palangi, H., et al.: Deep sentence embedding using long short-term memory networks: analysis and application to information retrieval. IEEE/ACM Trans. Audio Speech and Lang. Process. **24**(4), 694–707 (2016)
8. Tu, Z., Lu, Z., Liu, Y., Liu, X., Li, H.: Modeling coverage for neural machine translation. In: Proceedings of 40th Annual Meeting of the Association for Computational Linguistics, pp. 76–85 (2016)
9. Wang, X., Tu, Z., Zhang, M.: Incorporating statistical machine translation word knowledge into neural machine translation. IEEE/ACM Trans. Audio Speech and Lang. Process. **26**(12), 2255–2266 (2018)
10. Wang, X., Lu, Z., Tu, Z., Li, H., Xiong, D., Zhang, M.: Neural machine translation advised by statistical machine translation. In: Proceedings of AAAI Conference on Artificial Intelligence (2016)
11. Zhu, X., Yang, M., Zhao, T., Zhu, C.: Minimum Bayes-risk phrase table pruning for pivot-based machine translation in internet of things. IEEE Access **6**, 55754–55764 (2018)