KG-IQES: An Interpretable Quality Estimation System for Machine Translation Based on Knowledge Graph

Junhao Zhu[*] zhujunhao@huawei.com
Huawei Translation Center,Dongguan,China
Min Zhang[†] zhangmin186@huawei.com
Huawei Translation Center,Beijing,China
Hao Yang[‡] yanghao30@huawei.com
Huawei Translation Center,Beijing,China
Song Peng Huawei Translation Center,Dongguan,China
Zhanglin Wu Huawei Translation Center,Beijing,China
Yanfei Jiang Huawei Translation Center,Dongguan,China
Xijun Qiu Huawei Translation Center,Dongguan,China
Weiqiang Pan Huawei Translation Center,Dongguan,China
Ming Zhu Huawei Translation Center,Beijing,China
Miaomiao Ma Huawei Translation Center,Beijing,China
Weidong Zhang Huawei Translation Center,Beijing,China

Abstract

The widespread use of machine translation (MT) has driven the need for effective automatic quality estimation (AQE) methods. How to enhance the interpretability of MT output quality estimation is well worth exploring in the industry. From the perspective of the alignment of named entities (NEs) in the source and translated sentences, we construct a multilingual knowledge graph (KG) consisting of domain-specific NEs, and design a KG-based interpretable quality estimation (QE) system for machine translations (KG-IQES). KG-IQES effectively estimates the translation quality without relying on reference translations. Its effectiveness has been verified in our business scenarios.

1 Introduction

QE is an important task in MT research. It directly reflects the quality of MT output in real time during real-world application. Existing popular AQE metrics of MT output include: (1) lexicon-

[*]Equal contribution, shared first authorship
[†]Equal contribution, shared first authorship
[‡]Corresponding Author

Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track, pages 162–170
September 4–8, 2023, Macau SAR, China.

based BLEU (Papineni et al., 2002), which performs n-gram matching between the MT output and reference translations and calculates the matching score; (2) embedding-based BLEURT (Sellam et al., 2020) and BERTScore (Zhang* et al., 2020), which compare the semantic vector distance between the reference translation and MT output. However, both of the metrics need to rely on reference translations, which are scarce in real-world application (Freitag et al., 2022) and require heavy labor costs to construct. In recent years, various metrics have been proposed for reference-free QE. One of the current state-of-the-art metrics is model-based COMET-QE (Rei et al., 2021), which directly calculates the quality score by using the source sentence and the translation through neural network, but leads to poor interpretability. A more interpretable metric is Knowledge-Based Machine Translation Evaluation (KoBE) (Gekhman et al., 2020), a Google-proposed system-level metric based on KG for common domains, which predicts the translation quality by calculating the similarity between bilingual entities in the source and translated sentences and linked entities in the KG. Although this metric has good interpretability and good performance in system-level evaluations of common domains, it is ineffective in specific domains. This is because in specific domains, each paragraph or sentence contains a large number of domain-specific entities, and these entities can only be effectively covered by domain-specific KG, not KG for common domains.

To address these shortcomings, we design a highly interpretable segment-level QE system based on multilingual KG for the Wireless Network[1] domain, that is, an interpretable quality estimation system for machine translation based on knowledge graph (KG-IQES). KG-IQES identifies NEs in the source sentence first, queries the NE translations in the target language from the KG constructed offline, and matches the translated NEs in the KG with NEs in the translated sentence. In this way, the translation quality can be estimated, and the entity alignment details can be displayed (see example in Figure 1).

Overall, our contributions are as follows:

- We design an interpretable QE system for MT output (KG-IQES) and apply it to the Wireless Network domain.
- We propose an effective bilingual entity alignment method for KG-IQES. It achieves much better performance than Fast-Align.
- In the Wireless Network domain, we construct a large-scale multilingual KG, with more than 1.7 million nodes and more than 6 million edges.
- According to experiments in the Wireless Network domain, our system KG-IQES effectively identifies 44.15% of MT bad cases, outperforming KoBE in system-level evaluations.

2 Related Work

Reference-free QE of MT output is a task that predicts quality of the machine translations by scoring the source text and machine-translated text. Many metrics have been proposed in this regard, roughly divided into three categories:

1. Lexicon-based metrics: word-based evaluation metrics that usually use a vocabulary and word statistics. For example, Popović et al. (Popović et al., 2011) designed a bag-of-word translation model, which accumulates the possibility of word pairs aligned with the source text to evaluate the quality of the translation. Specia et al. (Specia et al., 2013) used language-agnostic linguistic features extracted from source texts and translations to

[1]It is the biggest domain of our translation business.

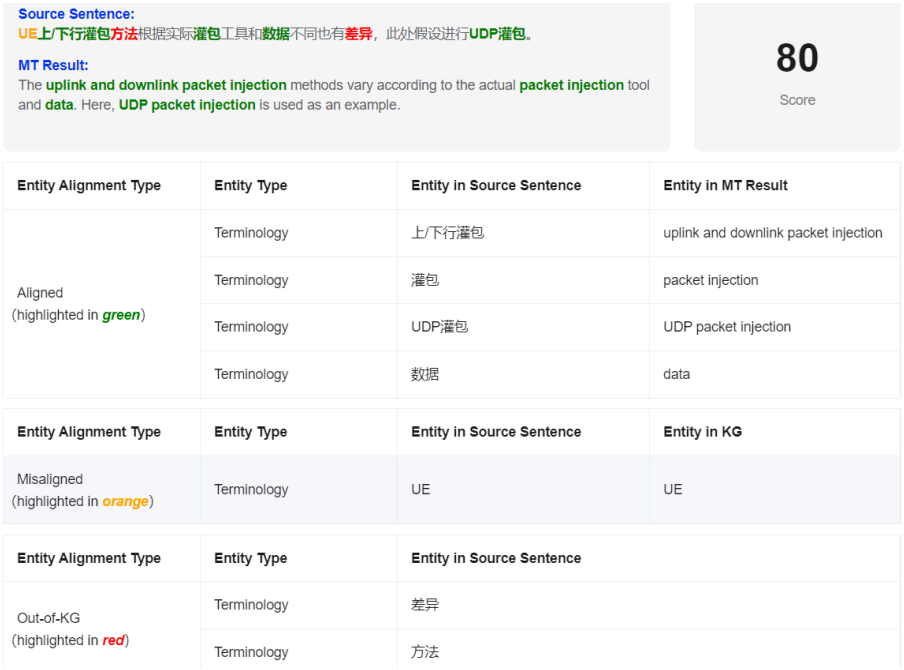

Figure 1: An example of KG-IQES that calculates the score of the MT result. KG-IQES highlights the aligned (green), misaligned (orange), and out-of-KG entities (red) in the source sentence and MT result, and directly shows the details.

estimate quality. KoBE, proposed by Google, calculates the quality score by calculating the recall of entities in the source text and the translation. These metrics are simple and effective, but are restricted by the coverage of aligned words.

2. Embedding-based metrics: word embedding-based evaluation metrics that use a pre-trained word embedding model to evaluate the semantic similarity of MT output. Examples include YiSi-2 (Lo and Larkin, 2020), SentSim proposed by Song et al. (Song et al., 2021), MoverScore metric (Zhao et al., 2019), and XMoverScore (Zhao et al., 2020) proposed by Zhao et al., which perform cross-language alignment directly, but word vectors trained by unsupervised methods during word embedding are usually noisy during their initialization.

3. Model-based metrics: these metrics that use a pre-trained deep neural network to evaluate the syntactic and semantic correctness of MT output. Classic metrics include COMET-QE and UniTE (Wan et al., 2022). The model-based metrics are highly effective when there is a large amount of high-quality manually annotated data. However, in most cases, such data is quite insufficient.

According to current research results, embedding-based and model-based metrics are more effective, but the interpretability is poor. They are not intuitive, and it is difficult to apply them in real world. In this paper, we follow the lexicon-based quality evaluation direction for MT output, and design a highly interpretable QE system KG-IQES. Its effectiveness is verified in the Wireless Network domain.

3 Proposed System

In this section, we provide a description of the KG-IQES system. As shown in Figure 2, the system consists of two subsystems: offline and online subsystems. The offline subsystem mainly constructs a multilingual KG based on the monolingual NER model and NE alignment model from the parallel corpora. The online subsystem provides an interpretable quality score for the MT output.

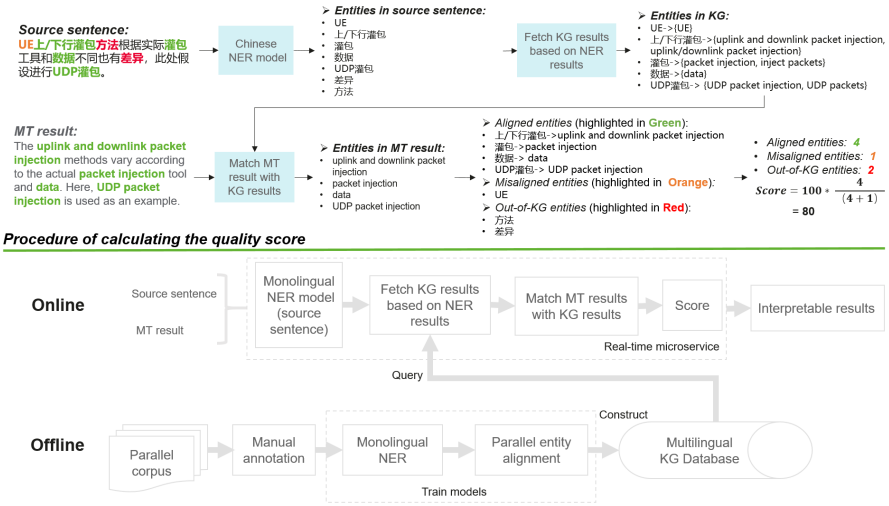

Figure 2: An overview of the KG-IQES system workflow: After offline training of the NER model and entity alignment model, the multilingual KG is constructed. The trained NER model service is deployed online, the KG is queried to obtain the entity translations, and the MT output quality is estimated based on whether the output contains the entity translations.

3.1 Offline Subsystem

The first step of KG-IQES system construction is NE annotation. We preferentially use manual annotation by domain experts who are also professional translators, and then adopt distant supervised annotation as a supplement when a corresponding knowledge base is available. In manual annotation by domain experts, we first obtain domain-specific monolingual corpora of Chinese or English, and then extract some monolingual sentences for manual annotation. Before annotation, we train all domain experts so as to unify the annotation standards. After all the domain experts complete annotating NEs in the monolingual sentences, they conduct cross-checks to reduce incorrect annotations. When a corresponding knowledge base is available, we extract some other monolingual sentences to adopt distant supervised annotation, that is, annotating NEs in these sentences that are matched to NEs in the knowledge base.

3.1.1 NER

To train the NER model, we use the W2NER (Li et al., 2022) model architecture, which consists of three components: encoder layer, convolution layer, and co-predictor layer. In order to improve the training effect of the W2NER model, we use domain-specific data to fine-tune the pre-trained language model in the encoder layer to make it capable of domain-specific encoding. After that, we apply the fast gradient method (FGM) (Miyato et al., 2017) for adversarial training to improve the robustness of the NER model.

3.1.2 Multilingual Domain-specific KG Construction

To construct a multilingual domain-specific KG, we use the cross-lingual NE alignment method to extract bilingual NE pairs from the domain-specific bilingual corpus. We compare the Fast-Align (Dyer et al., 2013) and XLM-RoBERTa (Li et al., 2021) alignment methods. The former needs to use the NER model to identify NEs in bilingual data, and use the Fast-Align tool to extract bilingual NE pairs where NEs in source and translated sentences are aligned. The latter requires a small amount of annotated data to train the alignment model and predicts the bilingual NE pairs in the bilingual data. We choose XLM-RoBERTa, which has a higher alignment accuracy, to build a multilingual domain-specific KG. Also, 30,000 manually annotated bilingual term pairs in the existing knowledge base are used to construct the training dataset of the alignment model, improving performance of the alignment model.

3.2 Online Subsystem

Reference-free QE of MT output refers to scoring the MT output based on the source text. Base on the NER model and the multilingual domain-specific KG, we design an interpretable reference-free QE system for MT output, namely KG-IQES. In the estimation process, we first use the NER model to identify NEs from the source sentence, and then search for bilingual NE pairs from the multilingual domain-specific KG. After that, we classify NEs in the source sentence as aligned NEs (NE count m: $\sum_{i=1}^{m} align_i$), misaligned NEs (NE count n: $\sum_{j=1}^{n} misalign_j$), and out-of-KG NEs (NE count: $count_{out\text{-}of\text{-}kg}$) according to whether the NE translations in the bilingual NE pairs appear in the machine-translated sentence. The translation score $Score$ is the quality score of the machine-translated sentence using KG-IQES. Finally, the score was converted into a 100-point representation.

The formula for calculating the MT output quality score is:

$$Score = 100 \times \max\{\frac{(\sum_{i=1}^{m} \alpha_i \times align_i) - \eta \times count_{out\text{-}of\text{-}kg}}{\sum_{i=1}^{m} \alpha_i \times align_i + \sum_{j=1}^{n} \alpha_j \times misalign_j}, 0\} \quad (1)$$

In the above formula, α indicates the weight of each entity category and η is penalty coefficient for entities that are not in KG. If the weight of each category is the same and the KG can cover most entities, we can set α to 1 and η to a number close to 0 in the above formula.

$$Score = 100 \times \frac{m}{m+n} \quad (2)$$

As shown in Figure 1, m (number of aligned NEs) is 4, n (number of misaligned NEs) is 1, p (number of out-of-KG NEs) is 2, and the $Score$ is $\frac{100\times4}{4+1}$ = 80. It should be pointed out that in our system, p is 0 in most cases because the out-of-KG NEs, if found, will be added to our KG database immediately.

4 Experiments

The following describes the detailed experiment process in the Chinese-English corpus of the Wireless Network domain. First, we carry out NER model training and fine-tuning experiment based on the Chinese corpus of the Wireless Network domain. We then compare two typical cross-lingual NE alignment methods, Fast-Align and XLM-RoBERTa, in the process of building the bilingual KG (Chinese and English) in the Wireless Network domain. Finally, we use our KG-IQES system and Google's KoBE (Gekhman et al., 2020) to estimate the MT output quality of different translation models, and compare the correlation between the two estimation results with direct assessment (DA).

4.1 Data Setup

Before detailed results of each experimental stage are stated, it is necessary to describe the data setup. As shown in Table 1, we collect 6,065,351 Chinese (zh) sentences and 3,523,402 Chinese-English (zh-en) bilingual sentences from the Wireless Network domain.

Corpus Type	Sentences
Wireless zh corpus	6,065,351
Wireless zh-en bilingual corpus	3,523,402

Table 1: Corpus statistics.

4.2 Results

4.2.1 Domain-specific KG

In order to train the Chinese NER and Chinese-English alignment models specific to the Wireless Network domain, we randomly select 1200 Chinese sentences in the Wireless Network domain, and have the contained NEs annotated by domain experts. We use 1000 sentences as the training set and the remaining 200 sentences as the test set. We also use the collected NEs in the Wireless Network domain for distant supervised annotation of 100,000 monolingual sentences in the Wireless Network domain. In the experiments, we compare the results of using only one of these two annotation methods and using them together on NER model training. As shown in Table 2, using them together gets a better training result.

NER Model	Chinese
Manual annotation	70.92
Manual annotation + Distant supervised annotation	73.75
Manual annotation + Distant supervised annotation + In-domain pre-training	80.33
Manual annotation + Distant supervised annotation + In-domain pre-training + FGM adversarial training	**81.80**

Table 2: F1 scores of different NER models on the test set.

To further improve the performance of the NER model on the test set, we not only pre-train the language model used in the W2NER model architecture (we use bert-chinese-base [2] for Chinese NER), but also use FGM for adversarial training in the NER model training phase. As shown in Table 2, in-domain pre-training gets a better result, and the use of FGM adversarial training further improves the NER model training effect.

In the experiment of using the cross-lingual NE alignment method to construct Chinese-English KG for the Wireless Network domain, we compare two alignment methods: Fast-Align and XLM-RoBERTa. When conducting the Fast-Align alignment experiment, we use the Chinese NER model and English NER model to extract NEs from bilingual data, and use the Fast-Align tool to extract bilingual NE pairs where the NEs in Chinese are aligned with their translations in English. When conducting the XLM-RoBERTa alignment experiment, we randomly

[2] https://huggingface.co/bert-base-chinese

select 40,000 bilingual sentences to annotate the bilingual NE pairs, and use 35,000 of them as the training set and the remaining 5,000 as the test set to train the XLM-RoBERTa alignment model. Finally, we use the XLM-RoBERTa alignment model to predict all bilingual sentences and extract bilingual NE pairs. As shown in Table 3, the alignment accuracy of the XLM-RoBERTa method is much higher, so we choose it in our system.

Alignment Method	**Accuracy**
Fast-Align	75.23
XLM-RoBERTa	**98.09**

Table 3: The accuracy of different cross-lingual NE alignment methods on the test set.

In the end, we construct a Chinese-English KG of the Wireless Network domain with 756,390 Chinese NEs, 958,798 English NEs (English NEs are top 5 translations of the Chinese NEs by occurrence frequency in the bilingual corpus), and 6,738,190 cross-lingual aligned NE pairs as shown in Table 4. In addition, we randomly select 200 Chinese entities, check the corresponding English entities, and find that the accuracy rate of top 1 translations is 96%.

Node	Chinese entities	756,390
	English entities	958,798
Edge	Aligned NE pairs	6,738,190
Top 1 Accuracy		**96%**

Table 4: Multilingual KG statistics.

4.2.2 KG-IQES vs. KoBE

After constructing the KG of the Wireless Network domain in Chinese and English, we also conduct a reference-free QE experiment. In the KoBE and KG-IQES approaches, we use the knowledge base of the Wireless Network domain and the same NER model. We randomly select 1000 bilingual sentences from the corpus as the test set, and use 11 translation systems, including Google, Youdao, Baidu, and our MT systems, for testing. The system-level Pearson correlation coefficients between KoBE and DA and between KG-IQES and DA are calculated. As shown in Table 5, our KG-IQES system has a higher correlation with DA scores.

QE Method	**Pearson Correlation Coefficient**
KoBE	65.99
KG-IQES	**84.42**

Table 5: System-level Pearson correlation coefficients between the results of different QE methods and DA.

4.2.3 KG-IQES Effectiveness Verification

To further verify the effectiveness of our system in real-world application, we analyze and verify 308 bad cases of the MT system in the Wireless Network domain from the previous 9 months this year. It is found that, except for overtranslation, as shown in Table 6, the KG-IQES system can solve 63.55% of bad cases in two types: undertranslation and mistranslation. The problem of entity overtranslation is not yet solved (see Figure 3). We plan to solve it by extracting NEs from the translation and comparing them with NEs in bilingual NE pairs.

Bad Case Type	Count	Solved	Rate
undertranslation	143	91	**63.64%**
mistranslation	71	45	**63.38%**
overtranslation	94	0	0.00%
Total	308	136	**44.15%**

Table 6: KG-IQES solves 44.15% of MT bad cases in the previous 9 months.

Figure 3: An example of overtranslation case: we do not extract NEs from the translation, so we do not find the entity 'F3T members' in the MT result.

5 Conclusion

We propose the KG-IQES, a simple, efficient, and interpretable system based on KG for estimating the quality of MT output without relying on reference translations. KG-IQES consists of the offline and online subsystems. The offline subsystem mainly constructs a multilingual KG based on the monolingual NER model and NE alignment model from the parallel corpora. The online subsystem provides an interpretable quality score for the MT output. Effectiveness of KG-IQES is verified by experiments in Huawei's Wireless Network domain. In the future, we will mitigate the problem of overtranslation and expand KG-IQES to more languages and more domains.

References

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Freitag, M., Rei, R., Mathur, N., Lo, C.-k., Stewart, C., Avramidis, E., Kocmi, T., Foster, G., Lavie, A., and Martins, A. F. (2022). Results of wmt22 metrics shared task: Stop using bleu–neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68.

Gekhman, Z., Aharoni, R., Beryozkin, G., Freitag, M., and Macherey, W. (2020). Kobe: Knowledge-based machine translation evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3200–3207.

Li, B., He, Y., and Xu, W. (2021). Cross-lingual named entity recognition using parallel corpus: A new approach using xlm-roberta alignment. *arXiv preprint arXiv:2101.11112*.

Li, J., Fei, H., Liu, J., Wu, S., Zhang, M., Teng, C., Ji, D., and Li, F. (2022). Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10965–10973.

Lo, C.-k. and Larkin, S. (2020). Machine translation reference-less evaluation using yisi-2 with bilingual mappings of massive multilingual language model. In *Proceedings of the Fifth Conference on Machine Translation*, pages 903–910.

Miyato, T., Dai, A. M., and Goodfellow, I. (2017). Adversarial training methods for semi-supervised text classification. In *International Conference on Learning Representations*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Popović, M., Vilar, D., Avramidis, E., and Burchardt, A. (2011). Evaluation without references: Ibm1 scores as evaluation metrics. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 99–103.

Rei, R., Farinha, A. C., Zerva, C., van Stigt, D., Stewart, C., Ramos, P., Glushkova, T., Martins, A. F., and Lavie, A. (2021). Are references really needed? unbabel-ist 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040.

Sellam, T., Das, D., and Parikh, A. (2020). Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Song, Y., Zhao, J., and Specia, L. (2021). Sentsim: Crosslingual semantic evaluation of machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3143–3156.

Specia, L., Shah, K., De Souza, J. G., and Cohn, T. (2013). Quest-a translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79–84.

Wan, Y., Liu, D., Yang, B., Zhang, H., Chen, B., Wong, D., and Chao, L. (2022). Unite: Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127.

Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K. Q., and Artzi, Y. (2020). Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zhao, W., Glavaš, G., Peyrard, M., Gao, Y., West, R., and Eger, S. (2020). On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1656–1671.

Zhao, W., Peyrard, M., Liu, F., Gao, Y., Meyer, C. M., and Eger, S. (2019). Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.