DSEA-KoBE: Quality Estimation Based on Distantly Supervised Entity Alignment

Ming Zhu, Min Zhang, Junhao Zhu, Yanqing Zhao, Zhanglin Wu, Hao Yang, Song Peng, Shimin Tao
{Zhuming47,zhangmin186,zhujunhao,zhaoyanqing,wuzhanglin,yanghao30,pengsong2,taoshimin}@huawei.com
Huawei Technologies Co., Ltd.
Beijing, China

ABSTRACT

The interpretability of quality estimation without reference has always been a challenging issue in the evaluation of machine translation. Google's KoBE [1] method, which evaluates translation quality based on the matching rate of bilingual entities, has provided strong explainability. However, this approach faces challenges such as varying granularity in bilingual entity recognition and low map coverage, which limit its practical applicability. To address these limitations, we propose a DSEA-KoBE method that replaces the KoBE's entity link module with the Distantly Supervised Entity Alignment (DSEA) module. This distantly supervised approach effectively mitigates the problems related to graph coverage and varying granularity in bilingual entity recognition. We conduct experiments using KoBE's public data for English to Chinese (en-zh) and Chinese to English (zh-en) translations, and the results demonstrate the effectiveness of our proposed method.

KEYWORDS

quality estimation; interpretability; KoBE; DSEA-KoBE; distantly supervised

ACM Reference Format:
Ming Zhu, Min Zhang, Junhao Zhu, Yanqing Zhao, Zhanglin Wu, Hao Yang, Song Peng, Shimin Tao. 2023. DSEA-KoBE: Quality Estimation Based on Distantly Supervised Entity Alignment. In *Proceedings of December 8–9, 2023 (International Joint Conference on Knowledge Graphs).* Tokyo, Japan, 5 pages.

1 INTRODUCTION

Machine translation quality evaluation is an important part of natural language processing, which can be mainly divided into reference-based quality assessment and reference-free quality estimation. There are three ways for reference-based quality assessment, including N-gram-based similarity, edit distance matrix-based, and word embedding-based methods. BLEU[2], CHRF[3], and other methods are based on N-gram. BLEU calculates the similarity by measuring the word-level N-gram overlap between machine translation and reference translation, while CHRF is a character-level N-gram method. TER[4], WER[5], and PER[5] methods are typical edit distance-based approaches, and the main difference among these three methods lies in the definition of "errors" and the considered types of operations. TER considers insertion, deletion, substitution, and shift operations, and WER considers insertion, deletion, and substitution; in contrast, PER only considers insertion and deletion. Word embedding-based methods include BERTScore[6], BLEURT[7], COMET[8], YiSi[9], etc. BERTScore, BLEURT, and YiSi analyze and align at the lexical level between reference translation and machine translation using pre-trained models to calculate the similarity, while COMET is an end-to-end scoring model trained with manually scored data.

Compared with quality assessment with reference, quality estimation without reference is more challenging. Currently, quality estimation without reference mainly uses CometKiwi[10], KoBE, KG-BERTScore[11] and other methods. KoBE is more interpretable than CometKiwi and KG-BERTScore, and is easier to apply. Methods such as CometKiwi and TransQuest[12] use bilingual translation data with Direct Assessment (DA) scores to directly train a scoring model in an end-to-end manner. Those methods are feasible but not explainable. In contrast, the KoBE method creatively converts the translation quality problem into the translation accuracy problem of entities. KoBE identifies bilingual entities, links them to the same multilingual entity library, and then determines whether the entity translation is accurate by comparing the IDs of the linked entities, which shows a high degree of explainability. However, we find that in existing entity recognition, there is no uniform granularity cross-language entity recognition dataset, so the granularity of entities varies with languages. As shown in Figure 1, different entity granularities result in differences in entity links. In addition, the coverage of the multi-language knowledge graph is not high. Those facts result in incorrect entity links. To address this issue, we propose a Distantly Supervised Entity Alignment (DSEA) model and apply it to the entity matching module of KoBE. Experiments show that our proposed DSEA-KoBE method can solve this problem and is highly correlated with human scores.

Note that our DSEA method can also be used to build a cross-language graph. The knowledge graph is a highly explainable source for KoBE; however, entity linking at different granularities also introduces semantic models with lower explainability. As our model does not involve entity linking, we believe that our approach is competitive with KoBE in terms of explainability.

Our main contributions include:

- We find two problems of KoBE: (1) Different languages have different granularities of entity recognition. (2) The entity coverage of its knowledge graph is low, and test corpora cannot be well covered.
- We propose DSEA-KoBE, which is a quality estimation method based on KoBE. DSEA-KoBE alleviates the problem of map

International Joint Conference on Knowledge Graphs , Tokyo, Japan,

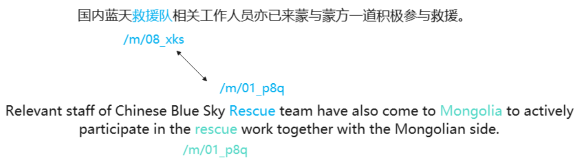
国内蓝天救援队相关工作人员亦已来蒙与蒙方一道积极参与救援。
/m/08_xks
/m/01_p8q
Relevant staff of Chinese Blue Sky Rescue team have also come to Mongolia to actively participate in the rescue work together with the Mongolian side.
/m/01_p8q

Figure 1: Problems caused by different entity granularities ***The real label of'救援队' is Rescue team.***

coverage and the problem of different granularity of entity recognition in different languages.

- We perform experiments on KoBE's zh-en and en-zh datasets and validate our method.

2 METHODS

Figure 2 shows the overall architecture of DSEA-KoBE. It consists of three modules: Named Entity Recognition (NER), DSEA, and score.

2.1 NER Module

NER is the basis of the entire DSEA-KoBE process, and all scores are based on the matching rate of entities. To better compare our method with KoBE, we directly use KoBE entity mentions and then perform entity alignment based on these entity mentions.

2.2 DSEA Module

Figure 3 shows the architecture of the DSEA module. In this module, we transform the entity alignment task into an NER task that extracts the corresponding entity in the parallel language for a given entity. We follow the basic model of CNN-Nested-NER[13], and change the input form for relation extraction.

We combine the source entities with the target sentences, and implement cross-language entity extraction using the multi-language capability of the XLM[14] pretrained model. For the alignment model, we use some bilingual entity pairs in the wiki and annotate parallel corpus through distant supervision.

2.3 Score Module

The scoring process is shown in Algorithm 1. First, we obtain the source entity through NER. Note that for better comparison with KoBE, we use the KoBE entity. We combine the source entity with the translation as the input of the alignment model, and then obtain the translation entity through the alignment model. Considering that the model based on distant supervision has some misjudgment, we use reverse verification technology, i.e., using the obtained translation entity and the source sentence as the input of the alignment model, to find the corresponding entity of the translation entity in the source sentence. We check whether the inferred entity is the same as the entity identified by NER (only the same text is considered, and the index information is not considered). If the two entities are the same, the matching is successful. The final score is obtained as follows:

$$F_{score} = \frac{match(src, mt)}{\sum_{k=1}^{n} count(entities_{s_k})}$$

Algorithm 1: DSEA-KoBE evaluation process

```
Input  : all source sentences $s_k \in S$ and machine
         translations $t_k \in T$ of $n$ sentence pairs
Output : a system-level score $F$
// $match(src, mt)$ the number of entities that
   are successfully matched.
$match(src, mt) = 0$ for each sentence pair $\{s_k,t_k\}$
  $\in \{S,T\}$ do
    // $entity_{s_k}$ indicates entities in the source
       sentence $s_k$, $entity_{s_{ki}}$ indicates the ith
       entity in the $entity_{s_k}$
    $entity_{st_{ki}} = Align(entity_{s_{ki}}, t_k)$
     $entity'_{s_{ki}} = Align(entity_{st_{ki}}, s_k)$
    if $entities'_{s_{ki}} == entities_{s_{ki}}$ then
        $match(src, mt) += 1$
    else
        do nothing
    end
end
// $count(entities_S)$ is the number of source
   entities
$F_{score} = \frac{match(src,mt)}{\sum_{k=1}^{n} count(entities_S)}$
```

3 EXPERIMENTAL SETUP

In order to comprehensively evaluate the effectiveness of our proposed method, we conduct experiments on two distinct tasks: entity alignment and system-level Quality Estimation (QE) sharing. For the entity alignment task, we employ two datasets to thoroughly assess the model's performance. The first dataset, referred to as the distant supervision dataset, is derived by mining 100,000 cross-language entity pairs from the Wikipedia corpus[1]. We extensively annotate this dataset using a large parallel corpus[2], obtaining a total of 50,000 instances. Out of these instances, 40,000 are designated as the training set, while the remaining 10,000 are allocated for development and testing purposes. The second dataset, denoted as KoBE-100, consists of a random sample of 100 data instances from KoBE's Chinese and English data. These instances are manually labeled, yielding a total of 734 instances with alignment relationships.

To evaluate the proposed DSEA method, we compare its performance against two baseline methods: Fast-Align[15] and Fast-Align (boundary)[16]. The Fast-Align method represents a standard alignment algorithm, while Fast-Align (boundary) incorporates the bilingual entity boundary information into the alignment process. Additionally, we compare our method against the entity linking method employed by KoBE itself, utilizing the KoBE-100 dataset.

Moving on to the system-level QE task, we conduct experiments using the zh-en and en-zh pairs extracted from the dataset provided by the WMT19 shared task. This dataset serves as the basis for

[1] https://dumps.wikimedia.org/
[2] https://www.statmt.org/wmt18/translation-task.htmldownload

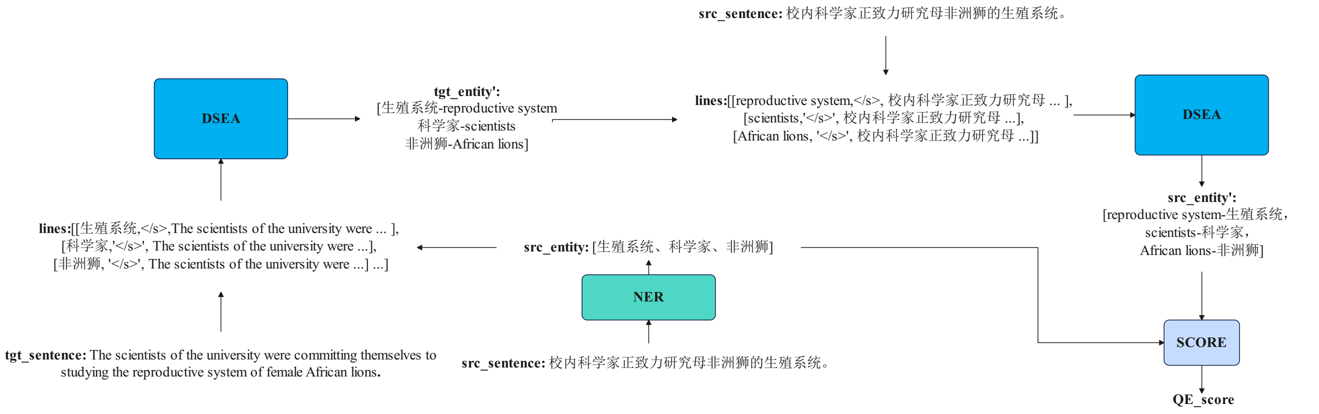
src_sentence: 校内科学家正致力研究母非洲狮的生殖系统。
DSEA
tgt_entity':
[生殖系统-reproductive system
科学家-scientists
非洲狮-African lions]
lines:[[reproductive system,</s>, 校内科学家正致力研究母 ...],
[scientists,'</s>', 校内科学家正致力研究母 ...],
[African lions, '</s>', 校内科学家正致力研究母 ...]]
DSEA
src_entity':
[reproductive system-生殖系统，
scientists-科学家，
African lions-非洲狮]
lines:[[生殖系统,</s>,The scientists of the university were ...],
[科学家,'</s>', The scientists of the university were ...],
[非洲狮, '</s>', The scientists of the university were ...] ...]
src_entity: [生殖系统、科学家、非洲狮]
NER
SCORE
tgt_sentence: The scientists of the university were committing themselves to studying the reproductive system of female African lions.
src_sentence: 校内科学家正致力研究母非洲狮的生殖系统。
QE_score

Figure 2: DSEA-KoBE

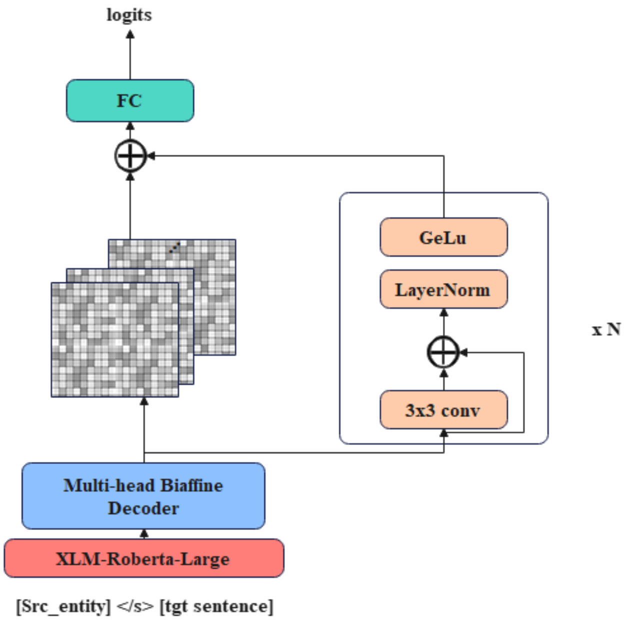

Figure 3: DSEA: Distantly Supervised Entity Alignment

evaluating the performance of our proposed method in the context of system-level QE.

Our model uses the CNN-Nested-NER architecture, and the pretraining model is XLM-R-Large. Five epochs are trained. The learning rate is 7e-6, cnn_dim is 200, and biaffine_size is 400. n_head is 4 and batch_size is 32.

method	Precision	Recall	F1
Fast-Align	0.613	0.568	0.590
Fast-Align(boundary)	0.765	0.651	0.703
DSEA	**0.961**	**0.961**	**0.961**

Table 1: Entity Alignment F1 on Distantly Supervised Dataset

method	Precision	Recall	F1
Fast-Align	0.660	0.375	0.478
Fast-Align(boundary)	0.782	0.451	0.572
KoBE	0.960	0.560	0.694
DSEA	0.923	**0.914**	**0.918**
DSEA(validation)	**0.961**	0.770	0.845

Table 2: Entity Alignment F1 on KoBE-100 dataset

Metric	en-zh	zh-en
BLEU	0.901	0.899
YiSi-2	-0.097	0.94
YiSi-2_srl	-0.118	**0.947**
BERTScore	-0.127	0.728
KoBE	0.216	0.907
KG-BERTScore	0.077	0.908
DSEA-KoBE	**0.375**	0.912
-(validation)	0.105	0.613

Table 3: System-level Pearson correlation with human judgements for zh-en and en-zh language pairs from the WMT19 metrics-without-references shared task. Best QE results are marked in bold.

4 RESULTS

Table 1 presents the results obtained on the distantly supervised dataset, indicating that the Fast-Align method achieves an F1 score of 0.590. We observe that the Fast-Align method may introduce entity alignment errors due to potential misalignment of subwords within the entities. To address this limitation, we propose an approach that incorporates the entity word boundaries in the target language as alignment constraints. This modification leads to a significant improvement, with the F1 score increased to 0.703. Furthermore, we evaluate our method on the manually labeled KoBE-100

src	mt	KoBE	DSEA-KoBE	Golden
未通过备案审核的班次不得招生培训	The classes that have not passed the archival filing and examination shall not enroll for training.	[班次, /m/0gz84v], [training, /m/014jg3], [archival, /m/01tygv]	[班次, classes]	[班次, classes]
腾讯新六大事业群	Tencent six new business groups.	[腾讯, /m/0403vtn], [事业群, /m/025w401], [Tencent, /m/0403vtn], [business groups, /m/0d06sy]	[腾讯, Tencent], [事业群, business groups]	[腾讯, Tencent], [事业群, business groups]

Table 4: Comparative example of matching results between KoBE and DSEA-KoBE

dataset, as depicted in Table 2. The entity link method employed by KoBE achieves a high precision of 0.960. Notably, precision appears to be more accurate than the recall in this case. However, the entity link method only achieves a recall of 0.560. Through analysis, we identify that the discrepancy in bilingual entity granularity is a key contributing factor to this problem, as exemplified in Figure 1. In contrast, the proposed DSEA method achieves higher precision and recall rates, benefiting from its ability to handle varying bilingual entity granularities.

To further improve entity alignment accuracy, we introduce a DSEA(validation) method based on reverse verification. This method leverages backtracking from the identified target entities, improving the alignment accuracy. DSEA(validation) achieves a precision of 0.961, surpassing the KoBE's entity-linking method, while considerably outperforming the entity-linking method in recall.

In our final set of experiments, we replace KoBE's entity link module with our DSEA and evaluate its performance on the zh-en and en-zh datasets of the reference-free system-level QE shared task from WMT19[17]. DSEA-KoBE surpasses KoBE in both zh-en and en-zh settings, as illustrated in Table 3. However, it is noteworthy that despite KoBE's lower recall rate, its method remains highly competitive in system-level QE. To further investigate this, we analyze KoBE's scoring mechanism and propose that the strong correlation between KoBE and DA scores in system-level QE relies on the high precision of KoBE's entity link methods. Even with a lower recall rate, the relative recall rate of machine translation from different systems tends to be consistent within the same domain data, thereby preserving the relative ranking in system-level scores. Consequently, we conduct additional experiments without reverse validation, confirming that the correlation between QE scores and DA scores is highly dependent on precision of entity matching. Although the recall rate has a minimal impact on the Pearson correlation of the DA score in system-level QE, it does influence the credibility of the score. When the number of alignment entities to be recalled is small, it becomes challenging to assess the reasonableness of the recall rate.

We present two examples in Table 4, showcasing KoBE's entity link example and our DSEA example. In Example 1, we observe that the term "班次" is identified as an entity in Chinese, whereas it is not recognized as an entity in English. This difference in entity granularity across languages accounts for the disparity. Instead, two additional entities are identified in English. In Example 2, we find that both the DSEA and KoBE's entity links successfully match "腾讯" and "Tencent" However, the entities "事业群" and "business group" are not included in the knowledge graph. Consequently, the entity link associates them with a similar entity, although the link's "id" reveals that the pair of entities does not match. Conversely, our proposed DSEA method aligns these two entities. These findings indicate the advantageous aspects of our approach compared to entity linking in KoBE.

5 CONCLUSION

In this paper, we analyze the shortcomings and advantages of the KoBE method, and propose a distantly supervised alignment method, which can avoid the difference of entity granularity among different entities to a certain extent, and mitigate the problem of low map coverage. Moreover, we analyze the effect of entity alignment accuracy and recall rate on QE scores, and verify our conjecture.

In our method, the position of entity alignment is not considered. In the future work, we will consider using supervised methods to implement entity alignment with position information.

REFERENCES

[1] Zorik Gekhman, Roee Aharoni, Genady Beryozkin, Markus Freitag, and Wolfgang Macherey. Kobe: Knowledge-based machine translation evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3200–3207, 2020.

[2] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002.

[3] Maja Popović. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, 2015.

[4] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, 2006.

[5] Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st annual meeting of the Association for Computational Linguistics*, pages 160–167, 2003.

[6] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2019.

[7] Thibault Sellam, Dipanjan Das, and Ankur Parikh. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, 2020.

[8] Craig Stewart, Ricardo Rei, Catarina Farinha, and Alon Lavie. Comet-deploying a new state-of-the-art mt evaluation metric in production. In *AMTA (2)*, pages 78–109, 2020.

[9] Chi-kiu Lo. Yisi-a unified semantic mt quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, 2019.

[10] Nuno M. Guerreiro Chrysoula Zerva Ana C. Farinha Christine Maroti José G. C. de Souza Taisiya Glushkova Duarte M. Alves Alon Lavie Luisa Coheur André F. T. Martins Ricardo Rei, Marcos Treviso. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 2022.

[11] Junhao Zhu Zhanglin Wu Hao Yang Song Peng Shimin Tao Ming Zhu, Min Zhang. Kg-bertscore: Incorporating knowledge graph into bertscore for reference-free machine translation evaluation. In *Proceedings of the 11th International Joint Conference on Knowledge Graphs*, 2022.

[12] Tharindu Ranasinghe, Constantin Orăsan, and Ruslan Mitkov. Transquest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5070–5081, 2020.

[13] Xiaonan Li Xipeng Qiu Hang Yan, Yu Sun. An embarrassingly easy but strong baseline for nested named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1442–1452, 2023.

[14] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 2020.

[15] NA Smith C Dyer, V Chahuneau. 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, 2013.

[16] Min Zhang, Song Peng, Hao Yang, Yanqing Zhao, Xiaosong Qiao, Junhao Zhu, Shimin Tao, Ying Qin, and Yanfei Jiang. Entityrank: Unsupervised mining of bilingual named entity pairs from parallel corpora for neural machine translation. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 3708–3713, 2022.

[17] Erick Fonseca, Lisa Yankovskaya, André FT Martins, Mark Fishel, and Christian Federmann. Findings of the wmt 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, 2019.