Knowledge Graphs Enhanced Neural Machine Translation

Yang Zhao[1,2], **Jiajun Zhang**[1,2], **Yu Zhou**[1,4] and **Chengqing Zong**[1,2,3]
[1]National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3]CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
[4]Beijing Fanyu Technology Co., Ltd, Beijing, China
{yang.zhao, jjzhang, yzhou, cqzong}@nlpr.ia.ac.cn

Abstract

Knowledge graphs (KGs) store much structured information on various entities, many of which are not covered by the parallel sentence pairs of neural machine translation (NMT). To improve the translation quality of these entities, in this paper we propose a novel KGs enhanced NMT method. Specifically, we first induce the new translation results of these entities by transforming the source and target KGs into a unified semantic space. We then generate adequate pseudo parallel sentence pairs that contain these induced entity pairs. Finally, NMT model is jointly trained by the original and pseudo sentence pairs. The extensive experiments on Chinese-to-English and English-to-Japanese translation tasks demonstrate that our method significantly outperforms the strong baseline models in translation quality, especially in handling the induced entities.

1 Introduction

Neural machine translation (NMT) based on the encoder-decoder architecture becomes a new state-of-the-art approach due to its distributed representation and end-to-end learning [Luong *et al.*, 2015; Vaswani *et al.*, 2017].

During translation, entities in a sentence play an important role, and their correct translation can heavily affect the whole translation quality of this sentence. Therefore, due to the importance of the entities, various methods are proposed to improve their translation [Zhang and Zong, 2016; Dinu *et al.*, 2019; Ugawa *et al.*, 2018; Wang *et al.*, 2019]. Among them, a kind of methods aim to incorporate the knowledge graphs (KGs) to improve the entity translation.

In many languages and domains, people construct various large-scale KGs to organize structured knowledge on entities. Meanwhile, some studies incorporate KGs into NMT to enhance the semantic representation of the entities in sentence pairs and improve the translation [Shi *et al.*, 2016; Lu *et al.*, 2018; Moussallem *et al.*, 2019]. However, these studies have a drawback that they only focus on the entities that both appear in KGs and training sentence pair dataset

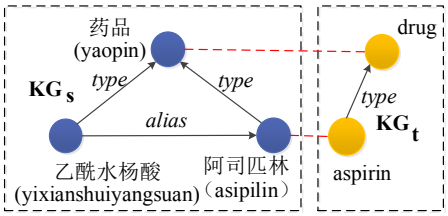

Figure 1: An example to show that the non-parallel KGs can also induce the translation results of $K-D$ entities. In the example two translation pairs can be extracted: "*asipilin-aspirin*" and "*yaopin-drug*" (shown in the red dotted line). Although the entity "*yixianshuiyangsuan*" is a $K-D$ entity, while it may be translated into "*aspirin*", since the source triple "*(asipilin, alias, yixianshuiyangsuan)*" indicates that "*yixianshuiyangsuan*" is another name for "*asipilin*".

(We denote these entities as $K+D$ entities[1]). Actually, besides these $K+D$ entities, KGs also contain many entities which do not appear in the training sentence pair dataset (We denote these entities as $K-D$ entities, whose formal definition can be found in Section 3). While these $K-D$ entities have been ignored in previous studies.

In this paper we think that these $K-D$ entities seriously harm the translation quality while KGs could alleviate this problem. Fig. 1 shows an example that assuming two translation pairs can be extracted from Chinese-to-English parallel sentence pairs, i.e., "*asipilin-aspirin*" and "*yaopin-drug*". Meanwhile, the source entity "*yixianshuiyangsuan*" is a $K-D$ entity and does not appear in the parallel sentence pairs. While we can induce that this entity may be translated into "*aspirin*", since the source triple "*(asipilin, alias, yixianshuiyangsuan)*" indicates that "*yixianshuiyangsuan*" is another name for "*asipilin*".

Therefore, in this paper we propose an effective method incorporating non-parallel source and target KGs into the NMT system. With the help of KGs, the proposed method could enable the NMT to learn new entity translation pairs containing the $K-D$ entities. More specifically, the proposed method contains three steps: 1) Bilingual $K-D$ entities induction: in this step we first extract the seed pairs from the phrase translation table. We then transform the source and target KGs into a unified semantic space by minimizing the distance be-

[1]K denotes KGs and D denotes the sentence pair dataset.

tween source and target entities in the seed pairs. We finally induce the translation results of the *K−D* entities under this semantic space. 2) Pseudo parallel sentence pairs generation: we generate adequate pseudo parallel sentence pairs containing the induced entity pairs. 3) Joint training: in this step we jointly train the NMT model by the original and pseudo senescent pairs, enabling NMT to learn the mapping between source and target entities in induced translation pairs. The extensive experiments on Chinese-to-English and English-to-Japanese translation tasks demonstrate that our method significantly outperforms the strong baseline models in translation quality, especially in handling the induced *K−D* entities.

We make the following contributions in this paper:

- We propose a method to incorporate the non-parallel KGs into NMT model.
- We design a novel approach to induce the translation results of the *K−D* entities with KGs, generate the pseudo parallel sentence pairs and promote NMT to make better predictions of *K−D* entities.

2 Background Knowledge

2.1 Neural Machine Translation

To date there are various NMT frameworks [Luong *et al.*, 2015; Vaswani *et al.*, 2017]. Among them, self-attention based framework (called as **Transformer**) achieves the state-of-the-art translation performance.

Transformer follows the encoder-decoder architecture, where the encoder transforms a source sentence X into a set of context vectors C. The decoder generates the target sentence Y from the context vectors C. Given a parallel sentence pair dataset $D = \{(X, Y)\}$, where X is the source sentence and Y is the target sentence, the loss function can be defined as:

$$L(D;\theta) = \sum_{(X,Y)\in D} log\ p(Y|X;\theta) \tag{1}$$

More details can be found in [Vaswani *et al.*, 2017].

2.2 Knowledge Embedding

The current KGs are always organized in the form of triples (h, r, t), where h and t indicate *head* and *tail* entities, and r denotes the relation between h and t, e.g., (*aspirin*, *type*, *drug*). Recently, various approaches are proposed to embed both entities and relations into a continuous low-dimensional space, such as TransE [Bordes *et al.*, 2013], TransH [Wang *et al.*, 2014] and TransR [Lin *et al.*, 2015]. Here we take TransE as an example to introduce the embedding methods.

TransE projects both relations and entities into the same continuous low-dimensional vector space $\mathbb{E}$. The goal of TransE is to make $\mathbb{E}(h) + \mathbb{E}(r) \approx \mathbb{E}(t)$. To achieve this, the score function is defined as:

$$f_r(h,t) = ||\mathbb{E}(h) + \mathbb{E}(r) - \mathbb{E}(t)|| \tag{2}$$

where $\mathbb{E}(h)$, $\mathbb{E}(r)$ and $\mathbb{E}(t)$ are the embeddings for h, r and t, respectively. $||.||$ means l_1 or l_2 norm. More details can be found in [Bordes *et al.*, 2013].

3 Problem Definition

In this paper we use the following three data resources to train a NMT model θ:

1) **Parallel Sentence Pairs** $D = \{(X, Y)\}$, where X denotes the source sentence. Y denotes the target sentence.

2) **Source KG** $KG_s = \{(h_s, r_s, t_s)\}$, where h_s, t_s and r_s denote the head entity, tail entity and relation in source language, respectively.

3) **Target KG** $KG_t = \{(h_t, r_t, t_t)\}$, where h_t, t_t and r_t denote the head entity, tail entity and relation in target language, respectively.

Since the parallel KGs are difficult to obtain, in this paper KG_s and KG_t are not parallel, Meanwhile, we assume that KG_s and KG_t contain many entities which do not appear in the parallel sentence pairs D. We called these entities as ***K−D* entities**. Formally, *K−D* entities set $\mathbb{O}$ can defined by

$$\begin{aligned} \mathbb{O}_{es} &= \{O_{es} | O_{es} \in KG_s \text{ and } O_{es} \notin D\} \\ \mathbb{O}_{et} &= \{O_{et} | O_{et} \in KG_t \text{ and } O_{et} \notin D\} \\ \mathbb{O} &= \mathbb{O}_{es} \cup \mathbb{O}_{et} \end{aligned} \tag{3}$$

where O_{es} and O_{es} denote the *K−D* source entity and target entity, respectively.

We think that although sentence pairs D may contain little translation knowledge on these *K−D* entities, the KGs could help to induce their translation results. Therefore, our goal in this paper is to improve the translation quality of these *K−D* entities with the help of KG_s and KG_t.

4 Method Descriptions

Fig. 2 shows the framework of our proposed method, which contains three steps: i) bilingual *K−D* entities induction, 2) pseudo sentence pairs generation and 3) joint training. Next we will introduce each step in the following each subsection.

4.1 Bilingual *K−D* Entities Induction

In this step we hope to induce the translation results of *K−D* entities. To achieve this goal, our main idea is to transform the source and target KGs into a unified semantic space, and then induce the translation results of these entities under this semantic space.

Specifically, **Algorithm 1** shows our bilingual *K−D* entities induction method, where the method first needs four preparations (line 1-4). We first represent KG_s and KG_t into the entity embedding $\mathbb{E}_s \in \mathbb{R}^{n\times d}$ and $\mathbb{E}_t \in \mathbb{R}^{m\times d}$, respectively (line 1-2). We then extract the phrase translation pairs $\mathbb{P} = \left\{(s, t, p_{(s,t)})\right\}$ from parallel sentence pairs D by statistical method[2], where s is the source phrase, t is the source phrase, $p_{(s,t)}$ is the translation probability (line 3). The last preparation is extracting *K−D* entity set $\mathbb{O}$ by Eq. (3). In the example of Fig. 2, there are three *K−D* entities "*yixian-shuiyangsuan*", "*purexitong*" and "*paracetamol*", where the first two are *K−D* source entities and the last one is *K−D* target entity.

With above preparations, we now need to construct the seed pair set $\mathbb{S}$ (line 5-8). If there is a phrase translation pair

[2]http://www.statmt.org/moses/

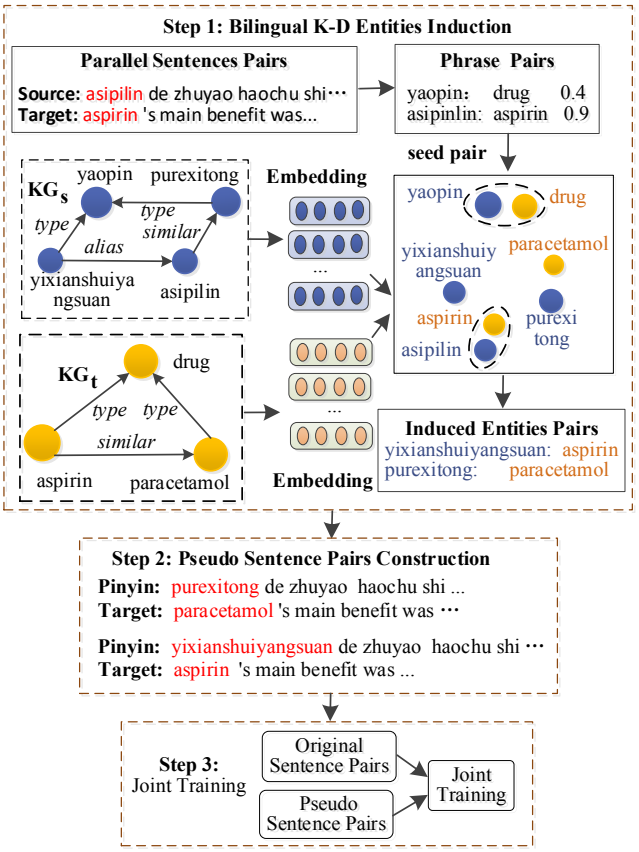

Figure 2: The proposed method which incorporates the non-parallel KGs into NMT.

$(s, t, p_{(s,t)})$ whose source phrase s belongs to KG_s and target phrase t belongs to KG_t, we add this phrase pair into seed pair set $\mathbb{S}$. In the example of Fig. 2, two phrase pairs "(*yaopin*, *drug*, 0.4)" and "(*asipilin*, *aspirin*, 0.9)" are selected into the seed pairs.

The derived $\mathbb{E}_s \in \mathbb{R}^{n\times d}$ and $\mathbb{E}_t \in \mathbb{R}^{m\times d}$ are learned separately, making them be in different semantic spaces. Now our task is to transform $\mathbb{E}_s$ and $\mathbb{E}_t$ into a unified semantic space. Inspired by [Zhu *et al.*, 2017], we conduct a linear transformation and make the source entities and target entities in seed pairs as close as possible. Specifically, given a seed pair $(s, t, p_{(s,t)})$, we define a transformation matrix W, so that $W\mathbb{E}_s(s) \approx \mathbb{E}_t(t)$. Futhur more, we take the translation probability $p_{(s,t)}$ into consideration. If a seed pair with a larger probability, this seed pair has a larger weigh in the loss function. Therefore, the loss function can be defined as Eq. (4) (line 9).

The final task is to induce the translation results of $K{-}D$ entities (line 10-17). Given a $K{-}D$ source entity $O_{es} \in \mathbb{O}$ (line 10), we traverse each target entity e_t in KG_t (line 11). If the distance between O_{es} and e_t is lower than the pre-defined threshold δ (line 12), we treat pair (O_{es}, e_t) as a new induced translation pair and add it into induction set $\mathbb{I}$ (line 13). Similarly, given a $K{-}D$ target entity $O_{et} \in \mathbb{O}$ (line 14), we traverse each source entity e_s in KG_s (line 15). We also add the pair (e_s, O_{et}) into induction set $\mathbb{I}$, if the distance between e_s and O_{et} is lower than the pre-define threshold δ (line 16-17). In the example of Fig. 2, we induced two new pairs: "(*yixianshuiyangsuan*, *aspirin*)" and "(*purexitong*, *paracetamol*)". Now the set $\mathbb{I}$ contains all new induced translation pairs.

Algorithm 1 Bilingual $K{-}D$ Entities Induction Method

Input:
Parallel sentence pairs D; source KG KG_s; target KG KG_t; pre-defined hyper-parameter δ
Output:
Bilingual $K{-}D$ entities induction set $\mathbb{I}$
Algorithm:

1: represent KG_s into embedding $\mathbb{E}_s \in \mathbb{R}^{m\times d}$
2: represent KG_t into embedding $\mathbb{E}_t \in \mathbb{R}^{n\times d}$
3: extract the phrase translation pairs $\mathbb{P} = \{(s, t, p_{(s,t)})\}$, where s is the source phrase, t is the source phrase, $p_{(s,t)}$ is the translation probability.
4: extract $K{-}D$ entity set $\mathbb{O}$ by Eq. (3)
5: initialize the seed set $\mathbb{S} = \{\}$
6: **for** each phrase pair $(s, t, p_{(s,t)}) \in \mathbb{P}$ **do**
7: **if** $s \in KG_s$ and $t \in KG_t$ **then**
8: add the phrase pair $(s, t, p_{(s,t)})$ into $\mathbb{S}$
9: learning the transform matrix W to represent $\mathbb{E}_s$ and $\mathbb{E}_t$ into a unified semantic space with seed set $\mathbb{S}$ by minimizing the following loss function

$$L = \sum_{(s,t,p_{(s,t)})\in\mathbb{S}} p_{(s,t)} ||W\mathbb{E}_s(s) - \mathbb{E}_t(t)|| \quad (4)$$

where $(s, t, p_{(s,t)})$ is the seed pair in $\mathbb{S}$. $\mathbb{E}_s(s)$ is the embedding for s and $\mathbb{E}_t(t)$ is the embedding for t.
10: **for** each $K{-}D$ source entity $O_{es} \in \mathbb{O}$ **do**
11: **for** each target entity $e_t \in KG_t$ **do**
12: **if** $||W\mathbb{E}_s(O_{es}) - \mathbb{E}_t(e_t)|| < \delta$ **then**
13: adding the induced pair (O_{es}, e_t) into $\mathbb{I}$
14: **for** each target $K{-}D$ entity $O_{et} \in \mathbb{O}$ **do**
15: **for** each source entities $e_s \in KG_s$ **do**
16: **if** $||W\mathbb{E}_s(e_s) - \mathbb{E}_t(O_{et})|| < \delta$ **then**
17: add the induced pair (e_s, O_{et}) into $\mathbb{I}$
return $\mathbb{I}$

4.2 Pseudo Sentence Pairs Generation

Now our goal is to generate the sentence pairs containing the induced entities pairs. The main idea is to transfer the context of seed pairs to the induced pairs which are close to this seed pairs. Specifically, if the distance between an induced pair $(i_s, i_t) \in \mathbb{I}$ and a seed pair $(s_s, s_t) \in \mathbb{S}$ is lower than a pre-defined hyper-parameter λ as follows:

$$||\mathbb{E}_s(i_s) - \mathbb{E}_s(s_s)|| + ||\mathbb{E}_t(i_t) - \mathbb{E}_t(s_t)|| < \lambda \quad (5)$$

we hope to transfer the context of seed pair (s_s, s_t) to that of induced pair (i_s, i_t). To achieve this goal, we first retrieve D and get all sentence pairs $\{(X_s, Y_s)\}$ containing the seed pair (s_s, s_t). Then we replace (s_s, t_s) in (X_s, Y_s) by the induced pair (i_s, i_t) and get the pseudo sentence pair (X_p, Y_p). Now the pseudo sentence pair (X_p, Y_p) contains the induced pair (i_s, i_t).

In the example of Fig. 2, assuming that both induced pairs "(*yixianshuiyangsuan*, *aspirin*)" and "(*purexitong*, *paracetamol*)" are close to the seed pair "(*asipilin*, *aspirin*)", we replace "(*asipilin*, *aspirin*)" by these two induced pairs and get the pseudo sentence pairs as shown in middle part of Fig. 2.

4.3 Joint Training

The final task is to train the NMT model θ with the original parallel sentence pairs D and pseudo parallel sentence pairs D_p. Our experiments (Section 6) show that the number of pseudo sentence pairs D_p is significantly less than that of original sentence pairs D. To overcome this imbalance problem, we first over-sample the pseudo sentence pairs D_p by n times and design the loss function by

$$L(\theta) = \sum_{(X,Y)\in D} log\ p(Y|X;\theta) + \sum_{(X_p,Y_p)\in D_p} log\ p(Y_p|X_p;\theta) \quad (6)$$

where the former one is the loss from the original data D, and the later one shows the loss from the over-sampled pseudo data D_p.

5 Experimental Setting

We test the proposed method on Chinese-to-English (CN⇒EN) and English-to-Japanese (EN⇒JA) translation tasks. The CN⇒EN parallel sentence pairs are extracted from LDC corpus, which contains 2.01M sentence pairs. On CN⇒EN task, we utilize three different KGs: i) **Medical KG**, where the source KG contains 0.38M triples[3] and the target KG contains 0.23M triples, which are filtered from YAGO [Suchanek *et al.*, 2007]. We construct 2000 medical sentence pairs as development set and 2000 medical sentence pairs as test set. ii) **Tourism KG**, where the source KG contains 0.16M triples. The target KG contains 0.28M triples, which are also filtered from YAGO[4]. We also construct 2000 sentence pairs on tourism as development set, and 2000 other sentence pairs as test set. iii) **General KG**, where the source KG is randomly selected from CN-DBpedia[5] and the target KG is randomly selected from YAGO. We choose the NIST 03 as development set and NIST 04-06 as test set. We use KFTT dataset as EN⇒JA parallel sentence pairs. The source and target KGs are DBP15K from [Sun *et al.*, 2017]. The statistics of training pairs and KGs are shown in Table 1.

We implement the NMT model based on the THUMT toolkit[6] and the knowledge embedding method based on the openKE toolkit[7]. We use the "base" version parameters of the Transformer model. On all translation tasks, we use the BPE [Sennrich *et al.*, 2016] method to merge 30K steps. We evaluate the final translation quality with case-insensitive BLEU for all translation tasks.

In this method, we compare the following NMT systems:

[3]http://www.openkg.cn/dataset/symptom-in-chinese

[4]The target KGs in Medical KG and Tourist KG are filtered by retaining the triples which contain the pre-defined key words.

[5]http://www.openkg.cn/dataset/cndbpedia

[6]https://github.com/THUNLP-MT/THUMT

[7]https://github.com/thunlp/OpenKE

Task	Pair	Knowledge Graph	Dev/Test
CH⇒EN	2.01M	Medical (0.38M/0.23M)	2000/2000
		Tourism (0.16M/0.28M)	2000/2000
		General (3.1M/2.5M)	919/3870
EN⇒JA	0.44M	DBP15k (0.16M/0.24M)	1166/1160

Table 1: The statistics of the training data. Column **Pair** shows the number of parallel sentence pairs. Column **Knowledge Graph** shows the name and number of triples (source/target). Column **Dev/Test** shows the number of sentences in development/test set.

1) RNMT: The baseline NMT system using two LSTM layers as encoder and decoder [Luong *et al.*, 2015].

2) Transformer: The state-of-the-art NMT system with self-attention mechanism.

3) Transformer+RC: This is the method which incorporates KGs by adding the *Relation Constraint* between the entities in the sentences [Lu *et al.*, 2018], whose goal is to get a better representation of *K+D* entities in sentence pairs.

4) Transformer/RNMT+KG: This is our proposed KGs enhanced NMT model on the basis of Transformer and RNMT, where we set the hyper-parameter δ (Algorithm 1) by 0.45 (Medical), 0.47 (Tourism), 0.39 (General) and 0.43 (DBP15K) and λ (Section 4.2) by 0.86 (Medical), 0.82 (Tourism), 0.73 (General) and 0.82 (DBP15K). The over-sample time n (Section 4.3) is set to 4 (Medical), 3 (Tourism), 2 (General) and 3 (DBP15K), respectively. All these hyper-parameters are fine-tuned in development set.

6 Experimental Results

6.1 Translation Results

Results on RNMT model. Table 2 lists the main translation results of CN⇒EN and EN⇒JA translation tasks. We first compare our method with RNMT. Comparing the row 1 and row 4-6, the proposed RNMT+KG can improve over RNMT on all test sets. Specifically, when utilizing the medical, tourism and general KG, the proposed method can exceed RNMT by 1.29 (12.54 vs. 11.25), 0.88 (12.77 vs. 11.89) and 0.55 (41.89 vs. 41.34) BLEU points, respectively. Meanwhile, on EN⇒JA translation task, the improvement can reach 0.48 BLEU points (27.91 vs. 27.43).

Results on Transformer model. We conduct experiments to evaluate proposed method on the basis of Transformer. As shown in row 2 and row 7-9, our method can also improve the translation quality on Transformer, where with the help of these three KGs, the improvements can reach 1.12 (15.69 vs. 14.57), 0.90 (14.88 vs. 13.98) and 0.51 (44.91 vs. 44.40) BLEU points, respectively. Besides, on EN⇒JA translation task, the proposed Transformer+KG can outperform Transformer by 0.60 BLEU points (30.10 vs. 29.50).

Results on different embedding methods. We are also interested the results when we utilize different knowledge embedding methods. Here, we test the following three knowledge embedding methods: TransE, TransD and TransR. From the results (row 4-9), we can see that on all tasks, these three knowledge embedding methods can achieve similar BLEU scores.

#	Model	CH⇒EN Medical dev	CH⇒EN Medical test	CH⇒EN Tourism dev	CH⇒EN Tourism test	CH⇒EN General dev	CH⇒EN General test	EN⇒JA DBP15k dev	EN⇒JA DBP15k test
	Baselines								
1	RNMT [Luong *et al.*, 2015]	12.23	11.25	12.94	11.89	43.96	41.34	25.47	27.43
2	Transformer [Vaswani *et al.*, 2017]	14.73	14.57	14.92	13.98	45.80	44.40	27.34	29.50
3	Transformer+RC [Lu *et al.*, 2018]	14.92	14.79	14.91	14.11	46.20	44.83	27.61	29.83
	Our method								
4	RNMT+KG (TransE)	13.66†	12.54†	13.88†	12.77†	44.68†	41.89*	25.84*	27.91*
5	RNMT+KG (TransH)	13.71†	12.37†	13.84†	12.84†	44.49*	41.56	26.12*	27.73
6	RNMT+KG (TransR)	13.58†	12.29†	13.79†	12.99†	44.54†	41.77*	25.88*	28.03*
7	Transformer+KG (TransE)	15.96†	15.69†	15.58*	14.88†	46.36*	44.91†	27.79*	30.10*
8	Transformer+KG (TransH)	16.09†	15.43†	15.77†	14.69†	46.48†	44.80†	28.01†	29.88*
9	Transformer+KG (TransR)	15.70†	15.54†	**15.81**†	**14.94**†	46.49†	44.80†	27.81*	30.17*
10	Transformer+RC+KG (TransE)	**16.10**†	**15.81**†	15.71†	14.91†	**46.76**†	**45.20**†	**28.18**†	**30.33**†

Table 2: The BLEU scores of different methods on CN⇒EN and EN⇒JA translation tasks. "*" indicates that the proposed system is statistically significant better ($p < 0.05$) than the baseline system and "†" indicates $p < 0.01$.

Transformer+RC vs. Our method. We also compare the proposed method with Transformer+RC (row 3). The results show that our proposed method (row 7-9) can outperform Transformer+RC by 0.90 (15.69 vs. 14.79), 0.77 (14.88 vs. 14.11), 0.08 (44.91 vs. 44.83) and 0.27 (30.10 vs. 29.83) BLEU points, respectively. The results show the advantage of proposed methods. More importantly, on the basis of Transformer+RC, our method (row 10) can further improve the translation quality, indicating that Transformer+RC still faces the problem of $K-D$ entities, and our method can alleviate this problem.

6.2 Effect of Hyper-paramters

In **Algorithm 1**, we set a pre-defined hyper-parameter δ to determine the bilingual induced pairs. Table 3 shows the BLEU scores with different δ (medical KG). We can see that the BLEU score is the largest when $\delta = 0.45$. When δ exceeds 0.45, the BLEU score (dev) decreases from 15.96 to 14.94 BLEU points.

Meanwhile we are also curious about the precision of induced bilingual $K-D$ entities. Thus we randomly select 300 induced bilingual entity translation pairs under different δ and analyze the correct ratio manually. The results are also reported in Table 3 (Column **Precison**). From the results, we can see that with the increasing of hyper-parameter δ, more $K-D$ entity translation pairs can be induced, while the precision decreases from 43.1% to 13.7%. The results show that it is necessary to strike a balance between the number of $K-D$ entity translation pairs and the precision of that.

δ	# Pair	BLEU dev	BLEU test	Precision
Baseline	0	14.73	14.57	–
0.40	2.2k	14.98	15.07	43.1%
0.42	5.9k	15.18	15.33	36.4%
0.45	13.8k	**15.96**	**15.69**	31.9%
0.47	20.5k	15.28	15.33	20.3%
0.50	41.7k	14.94	14.48	13.7%

Table 3: The BLEU scores with different δ. # **Pair** shows the number of induced $K-D$ bilingual entity pairs. **Precison** shows correct ratio of induced $K-D$ bilingual entity pairs.

In section 4.2, we set a pre-defined hyper-parameter λ to generate pseudo sentence pairs. Fig. 3 shows the results (medical KG), where x axis denotes hyper-parameter λ, y axis denotes the BLEU score of development and test set. The number in the figure denotes the number of pseudo sentence pairs. We can see from the results that with the increasing of hyper-parameter λ, more pseudo sentence pairs can be generated. The BLEU score (dev) becomes the largest when $\lambda = 0.86$. We think the reason is that when λ becomes too large, pseudo sentence pairs may contain more noises, which consequently harm the final translation quality.

6.3 Analysis of $K-D$ Entities

In this paper we aim at enhancing the $K-D$ entities of NMT with KGs. Thus, we also analyze the results on $K-D$ entities of proposed methods. Specifically, the analysis is conducted on sentence level and entity level.

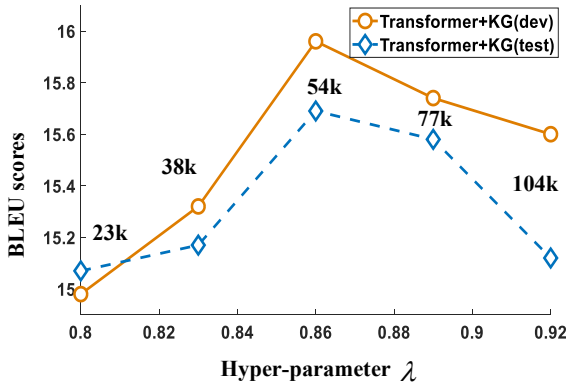

Figure 3: The effect of hyper-parameter λ, where x axis denotes hyper-parameter λ, y axis denotes the BLEU score of development and test set. The BLEU scores of development and test set become the largest when $\lambda = 0.86$.

Model	Sent w/o *K-D*	Sent w *K-D*
RNMT	13.25	9.96
RNMT+KG	13.54	11.75
Transformer	16.51	13.46
Transformer+KG	16.88	15.15

Table 4: The BLEU scores of sentences without *K-D* entities (Sent w/o *K-D*) and sentences with *K-D* entities (Sent w *K-D*).

On sentence level analysis, we divide the test sentences into two different parts: i) the sentence with $K-D$ entities (sent w $K-D$) and ii) the sentences without $K-D$ entities (sent w/o $K-D$). Table 4 reports results. From the results, we can see that our proposed method has little effect on the sentences without $K-D$ entities. While it can significantly improve the sentence with $K-D$ entities from 9.96 to 11.75 BLEU points (RNMT) and from 13.46 to 15.15 BLEU points (Transformer), respectively. The results indicate that our proposed method can produce better translation results on the sentences containing the $K-D$ entities.

We also analyze the results of induced $K-D$ entities on word level. Specifically, we randomly selected 300 sentences (medical KG), which contains 162 $K-D$ entities (267 times) and 72 $K+D$ entities (96 times). And we count the following three numbers: 1) correct ratio (times) of induced $K-D$ entities; 2) correct ratio (times) of un-induced $K-D$ entities; and 3) correct ratio (times) of $K+D$ entities. The statistics are reported in Table 5. From the result, we can see that Transformer+RC could improve the translation correct ratio (times) of $K+D$ entities from 36.5% (35) to 43.8% (42). And our method is most effective to induced $K-D$ entities, which improves the translation correct ratio (times) from 21.5% (31) to 31.3% (45). More importantly, when combining Transformer+RC and our method, RC+Ours can both improve the induced $K-D$ and $K+D$ to 31.9% (46) and 46.9% (45), respectively, which shows that our method and Transformer+RC are complementary.

Fig. 4 shows the example that the proposed method could improve the translation of induced $K-D$ entities. In the example, the mentioned $K-D$ entity "*yixianshuiyangsuan*" is totally translated into a wrong target phrase by Transformer. While the proposed Transformer+KG could overcome this mistake and produce the correct translation result "*asipirin*".

7 Related Work

In this paper we aim at incorporating the KGs into NMT to improve the $K-D$ entities. The related work can be divided into the following three categories :

Knowledge Graph in NMT. The early studies using the knowledge graph or semantic web are conducted in statistical machine translation framework [Moussallem *et al.*, 2018]. Recently, several studies incorporate the KGs into NMT, where [Shi *et al.*, 2016] proposes a knowledge-based semantic embedding for NMT by extracting the important semantic vectors with KGs. [Lu *et al.*, 2018] incorporates KGs by adding the relation constraint between the entities in the sentences. [Moussallem *et al.*, 2019] exploits the entity linking to disambiguate the entities found in a sentence. The biggest difference between our method and previous methods is that previous studies tend to enhance the semantic representation of $K+D$ entities in sentence pairs. While the goal of our method is to improve the translation of $K-D$ entities.

Model	*K−D*		*K+D*
	induced	un-induced	
Transformer	21.5% (31)	25.2% (31)	36.5% (35)
Transformer+RC	21.5% (31)	25.2% (31)	43.8% (42)
Ours	31.3% (45)	26.8% (33)	36.5% (35)
RC+Ours	**31.9% (46)**	**26.8% (33)**	**46.9% (45)**

Table 5: The correct ratio (times) on $K-D$ and $K+D$ entities.

Source: 不 建议 使用 乙酰水杨酸 和 布洛芬 。
Pinyin: bu jianyi shiyong yixianshuiyangsuan he buluofen 。
Target: aspirin and ibuprofen are not recommended .
NMT: it is not recommended to use the book of agriculture , yaoyang and braufen .
NMT+KG: it is not recommended to use aspirin and braufen .

Figure 4: An example to show that the proposed method could improve the induced $K-D$ entities.

Cross-lingual Knowledge Alignment. Our induced method is inspired by the work of knowledge alignment [Hao *et al.*, 2016; Zhu *et al.*, 2017; Chen *et al.*, 2017; Wang *et al.*, 2018; Cao *et al.*, 2019; Wu *et al.*, 2019]. The goal of knowledge alignment is to find entities in different KGs that refer to the same meaning. Different from these studies, our method aims to improve the translation quality of $K-D$ entities of NMT with KGs.

Incorporating Bilingual Lexicons or Phrases. Our method is also inspired by the studies of incorporating bilingual lexicons or Phrases into NMT [Zhang and Zong, 2016; Hasler *et al.*, 2018; Ugawa *et al.*, 2018; Zhao *et al.*, 2018a; Zhao *et al.*, 2018b; Dinu *et al.*, 2019; Huck *et al.*, 2019]. The difference between our method and these studies that is they utilize the external bilingual lexicons to improve the lexical translation, while we incorporate the KGs to improve the $K-D$ entities.

8 Conclusion

To address $K-D$ entities in NMT, we propose a knowledge graph enhanced NMT method. We first induce the translation results of the $K-D$ entities by utilizing non-parallel KGs, then generate pseudo parallel sentence pairs, and finally jointly train the NMT model. The extensive experiments on Chinese-to-English and English-to-Japanese tasks demonstrate that our method significantly outperforms the baseline models in translation quality, especially in handling the induced $K-D$ entities.

Acknowledgments

The research work described in this paper has been supported by the Natural Science Foundation of China under grant No. U1836221 and 61673380. The research work in this paper has also been Sponsored by CCF-Tencent Open Fund.

References

[Bordes *et al.*, 2013] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Proceedings of NeurIPS 2013*, pages 2787–2795, 2013.

[Cao *et al.*, 2019] Yixin Cao, Zhiyuan Liu, Chengjiang Li, Zhiyuan Liu, Juanzi Li, and Tat-Seng Chua. Multi-channel graph neural network for entity alignment. In *Proceedings of ACL 2019*, pages 1452–1461, 2019.

[Chen *et al.*, 2017] Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In *Proceedings of IJCAI 2017*, pages 1511–1517, 2017.

[Dinu *et al.*, 2019] Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. Training neural machine translation to apply terminology constraints. In *Proceedings of ACL 2019*, pages 3063–3068, 2019.

[Hao *et al.*, 2016] Yanchao Hao, Yuanzhe Zhang, Shizhu He, Kang Liu, and Jun Zhao. A joint embedding method for entity alignment of knowledge bases. In *Proceedings of CCKS 2016*, pages 3–14, 2016.

[Hasler *et al.*, 2018] Eva Hasler, Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne. Neural machine translation decoding with terminology constraints. In *Proceedings of NAACL-NLT 2018*, pages 506–512, 2018.

[Huck *et al.*, 2019] Matthias Huck, Viktor Hangya, and Alexander Fraser. Better OOV translation with bilingual terminology mining. In *Proceedings of ACL 2019*, pages 5809–5815, 2019.

[Lin *et al.*, 2015] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of AAAI 2015*, 2015.

[Lu *et al.*, 2018] Yu Lu, Jiajun Zhang, and Chengqing Zong. Exploiting knowledge graph in neural machine translation. In *Proceedings of CWMT 2018*, pages 27–38, 2018.

[Luong *et al.*, 2015] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of EMNLP 2015*, pages 1412–1421, 2015.

[Moussallem *et al.*, 2018] Diego Moussallem, Matthias Wauer, and Axel-Cyrille Ngonga Ngomo. Machine translation using semantic web technologies: A survey. *Journal of Web Semantics*, 51:1–19, 2018.

[Moussallem *et al.*, 2019] Diego Moussallem, Axel-Cyrille Ngonga Ngomo, Paul Buitelaar, and Mihael Arcan. Utilizing knowledge graphs for neural machine translation augmentation. In *Proceedings of K-CAP 2019*, pages 139–146, 2019.

[Sennrich *et al.*, 2016] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of ACL 2016*, pages 1715–1725, 2016.

[Shi *et al.*, 2016] Chen Shi, Shujie Liu, Shuo Ren, Shi Feng, Mu Li, Ming Zhou, Xu Sun, and Houfeng Wang. Knowledge-based semantic embedding for machine translation. In *Proceedings of ACL 2016*, pages 2245–2254, 2016.

[Suchanek *et al.*, 2007] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of WWW 2007*, pages 697–706, 2007.

[Sun *et al.*, 2017] Zequn Sun, Wei Hu, and Chengkai Li. Cross-lingual entity alignment via joint attribute-preserving embedding. In *Proceedings of ISWC 2017*, pages 628–644, 2017.

[Ugawa *et al.*, 2018] Arata Ugawa, Akihiro Tamura, Takashi Ninomiya, Hiroya Takamura, and Manabu Okumura. Neural machine translation incorporating named entity. In *Proceedings of COLING 2018*, pages 3240–3250, 2018.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of NeurIPS 2017*, pages 5998–6008, 2017.

[Wang *et al.*, 2014] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of AAAI 2014*, 2014.

[Wang *et al.*, 2018] Zhichun Wang, Qingsong Lv, Xiaohan Lan, and Yu Zhang. Cross-lingual knowledge graph alignment via graph convolutional networks. In *Proceedings of EMNLP 2018*, pages 349–357, 2018.

[Wang *et al.*, 2019] Tao Wang, Shaohui Kuang, Deyi Xiong, and António Branco. Merging external bilingual pairs into neural machine translation. *arXiv preprint arXiv:1912.00567*, 2019.

[Wu *et al.*, 2019] Yuting Wu, Xiao Liu, Yansong Feng, Zheng Wang, and Dongyan Zhao. Jointly learning entity and relation representations for entity alignment. In *Proceedings of EMNLP-IJCNLP 2019*, pages 240–249, 2019.

[Zhang and Zong, 2016] Jiajun Zhang and Chengqing Zong. Bridging neural machine translation and bilingual dictionaries. *arXiv preprint arXiv:1610.07272*, 2016.

[Zhao *et al.*, 2018a] Yang Zhao, Yining Wang, Jiajun Zhang, and Chengqing Zong. Phrase table as recommendation memory for neural machine translation. In *Proceedings of IJCAI 2018*, pages 4609–4615, 2018.

[Zhao *et al.*, 2018b] Yang Zhao, Jiajun Zhang, Zhongjun He, Chengqing Zong, and Hua Wu. Addressing troublesome words in neural machine translation. In *Proceedings of EMNLP 2018*, pages 391–400, 2018.

[Zhu *et al.*, 2017] Hao Zhu, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. Iterative entity alignment via joint knowledge embeddings. In *Proceedings of IJCAI 2017*, pages 4258–4264, 2017.