Knowledge Graph Enhanced Neural Machine Translation via Multi-task Learning on Sub-entity Granularity

Yang Zhao[1,2], **Lu Xiang**[1,2], **Junnan Zhu**[1,2], **Jiajun Zhang**[1,2],
Yu Zhou[1,2,3], and **Chengqing Zong**[1,2],
[1]National Laboratory of Pattern Recognition, Institute of Automation, CAS
[2]University of Chinese Academy of Sciences
[3]Beijing Fanyu Technology Co., Ltd
{yang.zhao,lu.xiang,junnan.zhu,jjzhang}@nlpr.ia.ac.cn
{yzhou,cqzong}@nlpr.ia.ac.cn

Abstract

Previous studies combining knowledge graph (KG) with neural machine translation (NMT) have two problems: i) Knowledge under-utilization: they only focus on the entities that appear in both KG and training sentence pairs, making much knowledge in KG unable to be fully utilized. ii) Granularity mismatch: the current KG methods utilize the entity as the basic granularity, while NMT utilizes the sub-word as the granularity, making the KG different to be utilized in NMT. To alleviate above problems, we propose a multi-task learning method on sub-entity granularity. Specifically, we first split the entities in KG and sentence pairs into sub-entity granularity by using joint BPE. Then we utilize the multi-task learning to combine the machine translation task and knowledge reasoning task. The extensive experiments on various translation tasks have demonstrated that our method significantly outperforms the baseline models in both translation quality and handling the 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; Gehring et al., 2017; Vaswani et al., 2017).

During translation, the translation quality of the entities in a sentence has a great influence on the translation quality of the whole sentence. However, translating these entities remains challenging (Moussallem et al., 2019) and various methods are proposed to improve the translation of entities (Zhang and Zong, 2016; Dinu et al., 2019; Ugawa et al., 2018; Wang et al., 2019). Among them, some methods aim at incorporating the knowledge graph (KG) to utilize their structured knowledge on entities and improve the entity translation. These studies utilize KG to enhance the semantic representing of entities in a sentence (Moussallem et al., 2019; Lu et al., 2018) or extract the important semantic vectors with KG (Shi et al., 2016). Although great efforts have been made to incorporate KG into NMT, we find the existing methods have the following two problems:

Knowledge Under-utilization: Given a KG (denoted by K) and a parallel sentence pair dataset (denoted by D), the full entity set U can be divided into four subsets as shown in Fig. 1 (a): 1) $\boldsymbol{K \cap D}$ **entities**, which appear in both K and D; 2) $\boldsymbol{D - K}$ **entities**, which only appear in D; 3) $\boldsymbol{K - D}$ **entities**, which only appear in K; 4) $\boldsymbol{U - (K \cup D)}$ **entities**, which are neither in K nor D. While previous studies (Shi et al., 2016; Moussallem et al., 2019; Lu et al., 2018) only focus on the $K \cap D$ entities, the other three subsets are ignored. Consequently, much knowledge on the other three subsets in KG is wasted.

Granularity Mismatch: The current KG methods, such as knowledge embedding methods (Bordes et al., 2013; Wang et al., 2014) and knowledge reasoning methods (Xiong et al., 2017), always utilize the entity as the basic granularity. While the NMT models utilize the sub-word (Sennrich et al., 2016) or hybrid word-character (Luong and Manning, 2016) as the translation granularity. This granularity mismatch between KG and NMT makes the current KG methods different to be utilized into NMT.

Proceedings of the 28th International Conference on Computational Linguistics, pages 4495–4505
Barcelona, Spain (Online), December 8-13, 2020

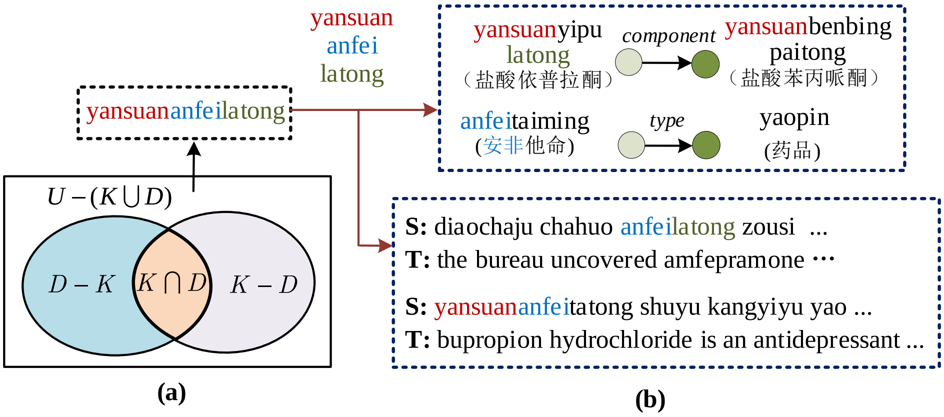


Figure 1: (a) The full entity set U can be divided into four different subsets. (b) shows an example that if we split a $U-(K\cup D)$ entity "`yansuananfeilatong`" into fine granularity "`yansuan`", "`anfei`", and "`latong`", each fine granularity component can be found in K and D.

Therefore, to address above two problems, we propose a multi-task learning method on sub-entity granularity to combine KG and NMT. Our main idea is to split the entities in K and D into fine granularity, and then enhance the semantic representation on this fine granularity. By doing so, we can achieve the following two goals: i) It can alleviate the knowledge under-utilization problem. Fig. 1 (b) gives an example, where source entity "`yansuananfeilatong`"[1] is $U-(K\cup D)$ entity, which is ignored by previous studies. While when we split this entity into fine granularity "`yansuan`", "`anfei`" , and "`latong`", we find that each fine granularity component can be found in both K and D. Therefore, we can enhance the semantic representation of these fine granularity components jointly with K and D. 2) The proposed method can also alleviate the granularity mismatch problem. It can enable both KG and NMT utilize an unified granularity (sub-entity), making knowledge in KG more easier to be utilized by the NMT.

Specifically, our proposed method contains two steps: i) **Joint BPE**, we utilize the Byte Pair Encoding (BPE) method (Sennrich et al., 2016) to jointly split the entities in K and D into sub-entity granularity. ii) **Multi-task learning**, we utilize multi-task learning to enhance the sub-entity embedding and the parameters in the neural model by machine translation task and knowledge reasoning task. In this step, we consider three different scenarios: a) only source KG is available; b) only target KG is available and c) both source and target KGs are available while they are non-parallel. The experimental results show that our proposed methods could improve the strong baseline models in translation quality, especially in handling the entities.

Generally, we make the following contributions in this paper:

1) We propose a KG enhanced NMT method to improve the entity translation under different scenarios: a) only source KG is available; b) only target KG is available and c) non-parallel source and target KGs are available.

2) We utilize the multi-task learning method on sub-entity granularity to make full use of semantic knowledge in KG and alleviate granularity mismatch problem.

2 Background Knowledge

2.1 Neural Machine Translation

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

Transformer contains two components: encoder θ_e and decoder θ_d, where encoder transforms a source

[1] It is the pinyin for Chinese entity "盐酸安非拉酮", whose English translation is "`amphetamine hydrochloride`".

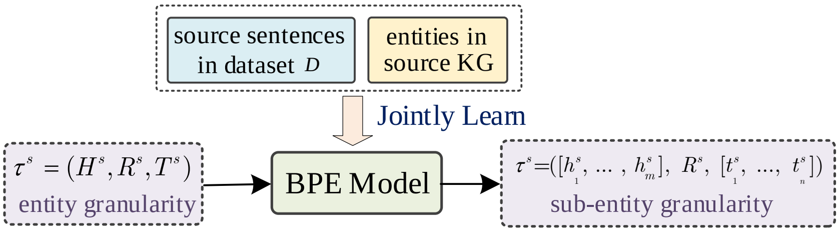

source sentences in dataset D
entities in source KG
Jointly Learn
$\tau^s = (H^s, R^s, T^s)$
entity granularity
BPE Model
$\tau^s = ([h^s_1, \dots, h^s_m], R^s, [t^s_1, \dots, t^s_n])$
sub-entity granularity

Figure 2: Joint BPE for source language.

sentence X into a set of context vectors C. The decoder can generate the target sentence Y from the context vectors C by calculating the predicted probability. Given a parallel sentence pair dataset $D = \{(X, Y)\}$, where X denotes the source sentence and Y denotes the target sentence, the loss function can be defined as:

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

Where θ_s and θ_d denote the parameters of encoder and decoder, respectively. More details can be found in (Vaswani et al., 2017).

2.2 Knowledge Graph

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., (`amphetamine hydrochloride`, `type`, `drug`). A large number of knowledge graphs (KGs) have been developed, such as YAGO (Suchanek et al., 2008) and Freebase (Bollacker et al., 2008). Meanwhile, various knowledge embedding methods (Bordes et al., 2013; Wang et al., 2014) and knowledge reasoning methods (Xiong et al., 2017) are proposed. These methods have been widely utilized by question answering systems(Bordes et al., 2015) and recommender systems (Zhang et al., 2016). While these methods utilize the entity as the granularity, making these methods different to be utilized into NMT models.

3 Problem Definition

We denote the parallel sentence pair dataset by $D = \{(X, Y)\}$, where X is the source sentence and Y is the target sentence. Besides D, we also need knowledge graph to improve the neural model. Considering that it is difficult to obtain the parallel KGs. Even in some low resource languages or domains, monolingual KG is also unavailable. Therefore, we consider the following three scenarios: a) *only source KG is available*, b) *only target KG is available*, and c) *both source and target KGs are available, while they are non-parallel*. In this paper, we aim to improve the NMT model in these three scenarios.

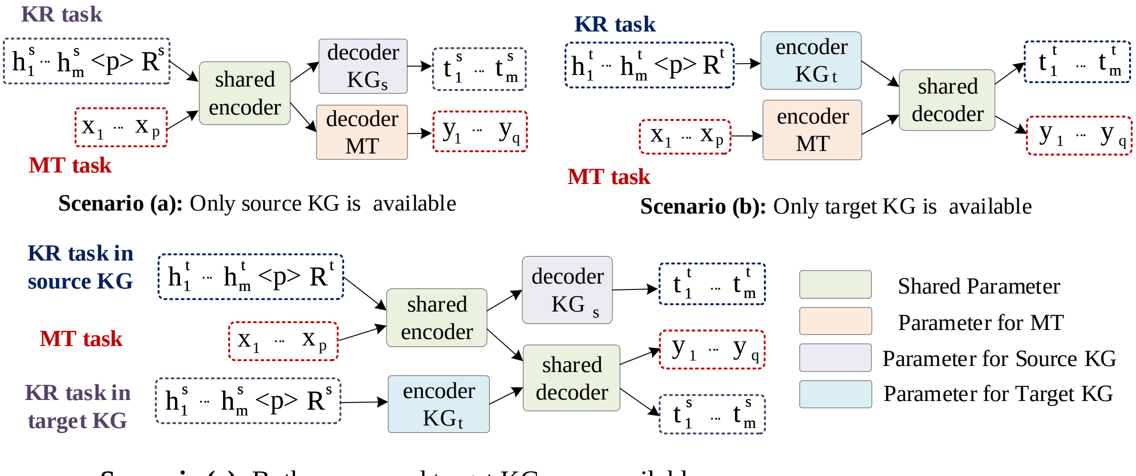
We denote the source KG by $K^s = \{(H^s, R^s, T^s)\}$, where H^s, T^s and R^s denote the head entity, tail entity and relation in source language, respectively. We denote the target KG by $K^t = \big\{(H^t, R^t, T^t)\big\}$, where H^t, T^t and R^t denote the head entity, tail entity and relation in target language, respectively.

4 Method Description

We propose a knowledge graph enhanced NMT via multi-task learning on sub-entity granularity, which contains two steps: i) joint BPE, and ii) multi-task learning. Next, we will introduce each step in the following subsections.

4.1 Joint BPE

In this step we aim to split the entities in K and D into the fine granularity. Considering the good performance of Byte Pair Encoding (BPE) method (Sennrich et al., 2016), we borrow this model to achieve our goal. Specifically, we first learn a joint BPE model with K and D. Then this joint BPE model can be utilized to split the entities in K and D into sub-entity granularity.



Scenario (a): Only source KG is available

Scenario (b): Only target KG is available

Scenario (c): Both source and target KGs are available

Figure 3: Multi-task learning between NMT task and KR task. In scenario (a), we share the encoder between NMT and KR of K_s. In scenario (b), we share the decoder between NMT and KR of K_t. In scenario (c), we share the encoder between NMT and KR of K_s and share the decoder between NMT and KR of K_t.

Fig. 2 shows an example of joint BPE for source language. By doing this, a source knowledge triple $\tau^s = (H^s, R^s, T^s)$ can be splited into $\tau^s = ([h_1^s, \dots, h_m^s], R^s, [t_1^s, \dots, t_n^s])$, where h_i^s and t_j^s denote the head and tail sub-entities in source sentence, respectively.

Similarly, if the target KG is available, a target knowledge triple $\tau^t = (H^t, R^t, T^t)$ can be splitted into $\tau^t = \left([h_1^t, \dots, h_p^t], R^t, [t_1^t, \dots, t_q^t]\right)$, where h_i^t and t_j^t denote the head and tail sub-entities in target sentence, respectively.

4.2 Multi-task Learning

Now we utilize the multi-task learning to improve the translation model with the KG. Specifically, there are two tasks:

- **Machine Translation Task:** predicting the target sentence $Y = \{y_1, ..., y_{|Y|}\}$, given the source sentence X. We can achieve this task by Transformer and calculate the predict probability $p(Y|X)$. Given the parallel sentence pair dataset $D = \{(X, Y)\}$, we can train the neural model by Eq. (1).

- **Knowledge Reasoning (KR) Task:** predicting the tail sub-entity sequence $T = [t_1, ..., t_n]$, given the head sub-entity sequence $H = [h_1, ..., h_m]$ and relation R. This task can also be seen as a sequence-to-sequence task, and thus be achieved by Transformer. Specifically, we first concatenate source head sub-entity sequence and relation by $\{h_1, ..., h_m, <p>, R\}$, where $<p>$ denotes the separator between head entity and relation. Then we utilize the Transformer framework to calculate the predicted probability $p(T|H, R)$. Given a knowledge graph $K = \{(H, R, T)\}$, we can train the neural model by

$$L(K; \theta_e, \theta_d) = \sum_{(H,R,T) \in K} \log p(T|H, R; \theta_e, \theta_d) \quad (2)$$

  where θ_e denotes the encoder and θ_d denotes the decoder.

We think the KR task can help the NMT model to learn better semantic representation. Therefore, we utilize multi-task learning to train the translation model with above two tasks. Fig. 3 shows our method.

Scenario (a): when only source KG is available, we share the encoder between NMT task and KR task to enhance the parameters of encoder (Fig. 3(a)). Therefore, we redesign loss function by

$$L(D,K^s;\theta_e^{\text{share}},\theta_d^{\text{mt}},\theta_d^{\text{kr}}) = \underbrace{\sum_{(X,Y)\in D}\log p(Y|X;\theta_e^{\text{share}},\theta_d^{\text{mt}})}_{\text{MT task}} + \underbrace{\sum_{(H^s,R^s,T^s)\in K^s}\log p(T^s|H^s,R^s;\theta_e^{\text{share}},\theta_d^{\text{kr}})}_{\text{source KR task}} \tag{3}$$

where θ_e^{share} denotes the shared encoder, θ_d^{mt} is the decoder for translation task, and θ_d^{kr} is the decoder for KR task.

Scenario (b): when only target KG is available, we share the decoder between these two tasks by

$$L(D,K^t;\theta_e^{\text{mt}},\theta_e^{kr},\theta_d^{\text{share}}) = \underbrace{\sum_{(X,Y)\in D}\log p(Y|,X;\theta_e^{\text{mt}},\theta_d^{\text{share}})}_{\text{MT task}} + \underbrace{\sum_{(H^t,R^t,T^t)\in K^t}\log p(T^t|H^t,R^t;\theta_e^{\text{kr}},\theta_d^{\text{share}})}_{\text{target KR task}} \tag{4}$$

where θ_e^{mt} denotes the encoder for NMT task, θ_e^{kr} denotes the encoder for KR task, and θ_d^{share} is the shared decoder.

Scenario (c): when both source and target KGs are available, we share the encoder between NMT and KR of K_s and share the decoder between NMT and KR of K_t. The loss function is redesigned by

$$\begin{aligned} L(D,K^s,K^t:\theta_e^{\text{share}},\theta_e^{\text{kr}},\theta_d^{\text{share}},\theta_d^{\text{kr}}) = & \underbrace{\sum_{(X,Y)\in D}\log p(Y|X;\theta_e^{\text{share}},\theta_d^{\text{share}})}_{\text{MT task}} \\ & + \underbrace{\sum_{(H^s,R^s,T^s)\in K^s}\log p(T^s|H^s,R^s;\theta_e^{\text{share}},\theta_d^{\text{kr}})}_{\text{source KR task}} \\ & + \underbrace{\sum_{(H^t,R^t,T^t)\in K^t}\log p(T^t|H^t,R^t;\theta_e^{\text{kr}},\theta_d^{\text{share}})}_{\text{target KR task}} \end{aligned} \tag{5}$$

where θ_e^{share} denotes the shared encoder, θ_e^{kr} denotes the encoder for target KR task, θ_d^{share} denotes the shared decoder, and θ_d^{kr} denotes the decoder for source KR task.

5 Experimental Setting

Dataset. We test the proposed method on Chinese-to-English (CN⇒EN), Chinese-to-Uygur (CN⇒UY) and Romanian-to-English (RO⇒EN) translation. 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) **General KG**, where the source KG is randomly selected from CN-DBpedia[2] and the target KG is randomly selected from YAGO. We choose the NIST 03 as development set and NIST 04-06 as test set. ii) **Medical KG**, where the source KG contains 0.41M triples[3] and the target KG contains 0.29M triples, which are filtered from YAGO[4]. We construct 2000 medical sentence pairs as development set and 2000 medical sentence pairs as test set. iii) **Tourism KG**, where the source KG contains 0.16M triples[5]. The target KG contains 0.28M triples, which are also filtered from YAGO. We also construct 2000 sentence pairs on tourism as development set, and 2000 other sentence pairs as test set. The CN⇒UY translation are extracted from CCMT-19 dataset. The RO⇒EN translation are extracted from TED dataset. The statistics of training pairs and KGs are shown in Table 1.

Training and Evaluation Details. We implement the NMT model based on the THUMT toolkit[6] (Zhang et al., 2017). 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 (Papineni et al., 2002) for all translation tasks.

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

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

[4] The target KG in medical and tourism KG is filtered by retaining the triples which contain the pre-defined key words.

[5] http://www.openkg.cn/dataset/tourist-attraction

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

Task	**Domain**	**Source KG**	**Target KG**	**Pair**	**Dev/Test**
CH⇒EN	General	3.3M	2.4M	2.01M	919/6146
	Medical	0.41M	0.29M		2000/2000
	Tourism	0.16M	0.28M		2000/2000
CH⇒UY	General	3.3M	-	0.22M	914/1678
RO⇒EN	General	-	2.4M	0.44M	1166/1160

Table 1: The statistics of the training data. Column **Pair** shows the number of parallel sentence pairs. Column **Source KG** and **Target KG** show the number of triples in source and target KGs. Column **Dev/Test** shows the number of sentences in development/test set.

#	Model		CH⇒EN		CH⇒UY	RO⇒EN
		General	Medical	Tourism		
1	Transformer	44.10	14.38	14.29	16.20	30.86
			D + source KG			
2	Transformer+RC	44.38	14.54	14.45	16.43	-
3	multi-task (entity)	44.49	14.72	14.43	16.52	-
4	multi-task (sub-entity)	44.72*	15.11*	14.78*	16.77*	-
			D + target KG			
5	Transformer+RC	44.29	14.47	14.31	-	31.05
6	multi-task (entity)	44.34	14.52	14.28	-	31.08
7	multi-task (sub-entity)	44.61	14.71*	14.63*	-	31.27
			D + source KG + target KG			
8	Transformer+RC	44.51	14.64	14.51	-	-
9	multi-task (entity)	44.53	14.89	14.47	-	-
10	multi-task (sub-entity)	44.89*	15.30†	14.91*	-	-

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

Comparing Methods. We compare the following NMT systems:

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

2) Transformer+RC: This is the method which incorporates KG by adding the *Relation Constraint* between the entities in the sentences (Lu et al., 2018), whose goal is to get a better representation of $K \cap D$ entities in sentence pairs.

3) multi-task(entity): This is multi-task learning on entity granularity.

4) multi-task(sub-entity): This is our proposed multi-task learning on sub-entity granularity.

6 Experimental Results

6.1 Translation Results with Different Scenarios

Table 2 lists the main translation results of CN⇒EN, CH⇒UY, and RO⇒EN translation tasks.

Results on source KG. When source KG is available, the BLEU scores are reported on line 2-4. From the results, we can see that the proposed multi-task(sub-entity) method can outperform the Transformer model and Transformer+RC method with source KG. Specifically, on CH⇒EN task, the proposed method can exceed the Transformer by 0.62 (44.72 vs. 44.10), 0.73 (15.11 vs. 14.38) and 0.49 (14.78 vs. 14.29) BLEU scores, respectively. Meanwhile, on CH⇒UY translation, the proposed method can also improve the neural model by 0.57 (16.77 vs. 16.20) BLEU points.

Results on target KG. Line 5-7 list the BLEU scores when target KG is available. From the results we can see that our model can also improve the translation results with target KG. For example, on CH⇒EN task, the proposed method can exceed the Transformer by 0.51 (44.61 vs. 44.10), 0.33 (14.71 vs. 14.38) and 0.34 (14.63 vs. 14.29) BLEU scores, respectively. Meanwhile, from results we can also find that in our method the source KG (line 2-4) can better improve the NMT model than target KG (line 5-7).

Results on source and target KGs. Line 8-10 list the BLEU scores when source and target KGs are available. From the results we can see that on CN⇒EN translation, our proposed methods can further improve the Transformer model to 44.89, 15.30 and 14.91 BLEU points, respectively. The results show that our proposed method is also effective on NMT when both source and target KGs are available.

Model	$K \cap D$	$K - D$	$D - K$	$U - (K \cup D)$
		CH⇒EN (General)		
Number	117	65	104	74
Transformer	64	38	51	27
Transformer+RC	68 (+3.42%)	39 (+1.54%)	50 (−0.96%)	27 (+0.00%)
multi-task (sub-entity)	68 (+3.42%)	43 (+7.69%)	51 (+0.00%)	30 (+4.05%)
		CH⇒EN (Medical)		
Number	80	89	73	96
Transformer	38	24	29	23
Transformer+RC	43 (+6.25%)	26 (+2.25%)	29 (+0.00%)	24 (+1.04%)
multi-task (sub-entity)	42 (+5.00%)	32 (+8.99%)	31 (+2.74%)	28 (+5.21%)

Table 3: The statistics and correct translation ratio of $K \cap D$, $K - D$, $D - K$ and $U - (K \cup D)$ entities. **Number** shows the numbers of each entity in all entities. **Transformer** shows the correct number of each entity in Transformer model. **Transformer+RC** and **multi-task (sub-entity)** show the correct number and improvement ratio of these entities in Transformer+RC method and our proposed model.

Model	Sentences w corrected entities	Sentences w uncorrected entities
Transformer	15.01	14.31
multi-task (sub-entity)	16.35	14.79

Table 4: The BLEU scores for sentences with corrected entities and sentences without corrected entities.

Source: 盐酸 安非 他酮 缓释片 的 不良 反应 有 ： 激动、失眠 和 头晕 等

Pinyin: yansuan anfei latong huanshipian de buliang fanying you: jidong , shimian he touyun deng

Reference: the reactions of amphetamine hydrochloride sustained release tablets are as follows: excitement, insomnia and dizziness

NMT: bad reaction of hydrochloric acid heads

Multi-task (sub-entity): bad reaction of hydrochloric acid is purgatory, insomnia and dizziness

Figure 4: An example to show that even through the proposed multi-task(sub-entity) cannot rectify the translation error of entity `amphetamine hydrochloride`, while it can also help to improve the translation of the whole sentence.

6.2 Analysis on Different Entities

Results on $K \cap D$, $K - D$, $D - K$ and $U - (K \cup D)$ entities. In the proposed method, we aim to improve the entity translation of NMT model. Thus we also analyze the results on different entities. As mentioned in introduction, the entities in test set can be divided into four subsets: 1) $K \cap D$, 2) $K - D$, 3) $D - K$ and 4) $U - (K \cup D)$. To analyze the results on these four subsets, we first randomly select 500 sentences on CH⇒EN (General) task and CH⇒EN (Medical) task, and then manually analyze the correct ratio of these four different subsets under three different methods (Transformer, Transformer+RC and multi-task (sub-entity)). Table 3 reports the results, where **Number** shows the numbers of each subset entity in all entities. From the result, we can reach the two following conclusions:

1) Our statistical results show that $K \cap D$ entities are only a part of full entity set U. Specifically, in general and medical domains, $K \cap D$ entities account for 32.5% and 32.7%, respectively.

2) The results show that our method can improve the correct ratio of $K \cap D$, $K - D$, $U - (K \cup D)$ entities. While our method slightly affects on $D - K$ entities. Transformer+RC method can only improve the translation quality of $K \cap D$ entity.

We also find an interesting phenomenon that even through the proposed multi-task(sub-entity) cannot rectify the translation error of entity itself, while it can also help to improve the translation of the whole sentence. Specifically, in above 500 sentences, we analyze the following two kinds of sentences: 1) the sentences that Transformer mistakenly translates an entity in it while our method rectifies this error (denoted by **sentences w corrected entities**); 2) the sentences that Transformer mistakenly translates an entity in it and our method does not rectify this error (denoted by **sentences w uncorrected entities**). The results are reported in Table 4. The results show that on sentences with corrected entities, our method can improve the translation quality from 15.01 to 16.35 BLEU points. On the sentences with uncorrected

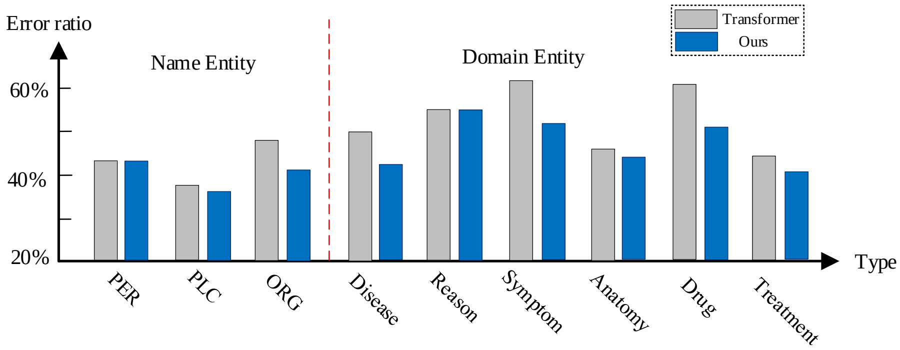


Figure 5: The translation error ratio of different name entities and domain entities.

Model	CH⇒EN (General)		CH⇒EN (Medical)	
	dev	test	dev	test
Transformer (sub-entity)	**44.80**	**44.10**	14.54	**14.38**
Transformer (word+character)	44.67	44.02	**14.69**	14.35
Transformer (character)	44.45	44.00	14.27	14.20
multi-task(sub-entity)	**45.56**	**44.89**	**15.94**	15.30
multi-task(word+character)	45.36	44.76	15.72	**15.39**
multi-task(character)	45.25	44.51	15.60	15.11

Table 5: The comparison of different granularities: sub-entity granularity, hybrid word-character granularity and character granularity.

entities, our method can still improve the translation quality from 14.31 to 14.79 BLEU points. We think the reason may be that our method can improve the semantic representation of sub-entity, which also help to improve the translation of the whole sentence. Fig. 4 shows an example, where Transformer mistakenly translates the entity `amphetamine hydrochloride`. Although the proposed method does not rectify this error, it improves the translation quality of the whole sentence.

Results on name entities and domain entities. Meanwhile, we also analyze the results on different name entities and domain entities. This analysis is conduced on general domain and medical domain of CN⇒EN. Specifically, we categorize the name entities by *PER*, *PLC* and *ORG*, and categorize the domain entities by *Disease*, *Reason*, *Symptom*, *Anatomy*, *Drug*, and *Treatment*. We also randomly 500 sentences and manually analyze the error ratio of these entities. Fig. 5 reports the results. From the results, we can reach the following two conclusions:

1) Our proposed method can reduce the translation error ratio of *ORG*, *Disease*, *Symptom*, *Drug*, and *Treatment*.

2) We also find that our method only slightly affects on *PER*, *PLC*, *Reason*, and *Anatomy*. We think there may be two reasons to cause this phenomenon: i) During translating these entities, KG is unnecessary[7]. ii) The KG we utilize in this paper does not cover the useful knowledge and semantic information which can improve the translation of these entities.

6.3 Comparison on Different Granularities

In this paper, we split the entities into sub-entity granularity. Actually, besides the sub-entity granularity, we also evaluate the other fine granularities: hybrid word-character granularity (Luong and Manning, 2016) and character granularity (Chung et al., 2016). The results are reported in Table 5. The results show that in both Transformer and our proposed multi-task method, the sub-entity granularity can produce better results than hybrid word-character granularity and character granularity.

[7] Take the entity *PER* as an example, assuming that the neural model tends to translate a person's name, while the KG always contains knowledge on his/her occupation, age or education, etc. Intuitively, this knowledge is not benefit for the translation of a person's name.

Model	Appeared Head Entities		Unseen Head Entities	
	Hit@10	Hit@20	Hit@10	Hit@20
TransE	23.4	30.3	-	-
TransH	23.6	30.7	-	-
Transformer(sub-entity)	23.2	30.5	17.9	23.3
Transformer(sub-entity)+MT	23.3	30.0	18.2	23.8

Table 6: The tail predicting results of knowledge reasoning task.

6.4 Results on KR Task

In this paper, we treat the knowledge reasoning as a sequence-to-sequence task and utilize this task to improve the performance of machine translation. Besides that, we are also curious about the following two questions: *1) how does it perform when we utilize the seq2seq model for KR task? 2) Can MT task also improve the KR task with multi-task learning?*

To answer above questions, we also evaluate the performance of KR task by 1) predicting the tail entity given an appeared head entity and a relation, and 2) predicting the tail entity given an unseen head entity and a relation. We utilize the Hit@10 and Hit@20 as the metrics. The two metrics calculate the ratio of predicted entity which ranks in the top 10 and top 20 in all entities. Table 6 lists the results, where TransE (Bordes et al., 2013) and TransH (Wang et al., 2014) are knowledge embedding methods for KR task. Transformer(sub-entity) is the seq2seq model on sub-entity granularity. Transformer(sub-entity)+MT is the multi-task learning with KR task and MT task.

Results given appeared head entities. In the scenario of appeared head entities, we can see that the seq2seq model on sub-entity granularity can achieve comparable results with TransE and TransH model. We can also see that in this scenario the MT task has little effect on KR task in multi-task learning.

Results given unseen head entities. In the scenario of unseen head entities, comparing to the TransE and TransH model, the seq2seq model has an advantage that it can handle the unseen head entities. In this scenario, we can see that MT task can slightly improve the performance of KR task.

7 Related Work

The related work can be divided into two categories and we describe each of them as follows:

Knowledge Graph in NMT. Moussallem et al. (2018) summarize the early studies using the knowledge graph or semantic web in statistical machine translation framework. Recently, several studies incorporate the KG into NMT, where Shi et al. (2016) propose a knowledge-based semantic embedding for NMT by extracting the important semantic vectors with KG. Lu et al. (2018) incorporate KG by adding the relation constraint between the entities in the sentences. Moussallem et al. (2019) exploit the entity linking to disambiguate the entities found in a sentence. While these studies only focus on the $\boldsymbol{K} \cap \boldsymbol{D}$ entities. Recently, Zhao et al. (2020) utilize the entity alignment methods to improve the $\boldsymbol{D} - \boldsymbol{K}$ entities. Different from these methods, the proposed methods utilize multi-task learning on sub-entity granularity to make full use of KG and improve the entity translation.

Incorporating bilingual lexicons and Phrases into NMT. Our method is also inspired by the studies of incorporating bilingual lexicons and phrases into NMT (Arthur et al., 2016; Zhang and Zong, 2016; Feng et al., 2017; Hasler et al., 2018; Zhao et al., 2018b; Zhao et al., 2018a; Dinu et al., 2019; Huck et al., 2019; Liu et al., 2019; Susanto et al., 2020). They utilize the external bilingual lexicons and phrases to improve the lexical and phrases translation. Different from these studies, we incorporate the KG to improve the entity translation.

8 Conclusion

To improve the entity translation and make1 full use of KG, in this paper we propose a KG enhanced NMT method with multi-task learning on sub-entity granularity. We first represent the entity in KG and parallel sentence pairs into sub-entity granularity. Then we utilize multi-task learning to improve the semantic represent of sub-entity and parameters in encoder or decoder. The proposed method can be utilized in different scenarios that source or/and target KG are available. The extensive experiments on various tasks demonstrate that our method significantly outperforms the baseline models in translation quality, especially in improving the entity translation.

Acknowledgments

The research work has been funded by the Natural Science Foundation of China under Grant No. U1836221 and 61673380. The research work in this paper has also been supported by Beijing Advanced Innovation Center for Language Resources and Beijing Academy of Artificial Intelligence (BAAI2019QN0504).

References

Philip Arthur, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of EMNLP 2016*, pages 1557–1567.

Kurt D. Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of SIGMOD 2008*.

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

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. *arXiv preprint arXiv:1506.02075*.

Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. 2016. A character-level decoder without explicit segmentation for neural machine translation. In *Proceedings of ACL 2016*, pages 1693–1703.

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

Yang Feng, Shiyue Zhang, Andi Zhang, Dong Wang, and Andrew Abel. 2017. Memory-augmented neural machine translation. In *Proceedings of EMNLP 2017*, pages 1390–1399.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of ICML 2017*, pages 1243–1252.

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

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

Xuebo Liu, Derek F Wong, Yang Liu, Lidia S Chao, Tong Xiao, and Jingbo Zhu. 2019. Shared-private bilingual word embeddings for neural machine translation. In *Proceedings of ACL 2017*, pages 3613–3622.

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

Minh-Thang Luong and Christopher D. Manning. 2016. Achieving open vocabulary neural machine translation with hybrid word-character models. In *Proceedings of ACL 2016*, pages 1054–1063.

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

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

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, pages 311–318.

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

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

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2008. Yago: A large ontology from wikipedia and wordnet. *Journal of Web Semantics*, 6(3):203 – 217.

Raymond Hendy Susanto, Shamil Chollampatt, and Liling Tan. 2020. Lexically constrained neural machine translation with levenshtein transformer. *arXiv preprint arXiv:2004.12681*.

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

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

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

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

Wenhan Xiong, Thien Hoang, and William Yang Wang. 2017. DeepPath: A reinforcement learning method for knowledge graph reasoning. In *Proceedings of EMNLP 2017*, pages 564–573.

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

Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, and Wei-Ying Ma. 2016. Collaborative knowledge base embedding for recommender systems. In *Proceedings of SIGKDD 2016*, pages 353–362.

Jiacheng Zhang, Yanzhuo Ding, Shiqi Shen, Yong Cheng, Maosong Sun, Huanbo Luan, and Yang Liu. 2017. Thumt: An open source toolkit for neural machine translation. *arXiv preprint arXiv:1706.06415*.

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

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

Yang Zhao, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2020. Knowledge graphs enhanced neural machine translation. In *Proceedings of IJCAI 2020*, pages 4039–4045.