Exploiting Knowledge Graph in Neural Machine Translation

Yu Lu[1,2], Jiajun Zhang[1,2], and Chengqing Zong[1,2,3]

[1] National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3] CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
{yu.lu,jjzhang,cqzong}@nlpr.ia.ac.cn

Abstract. Neural machine translation (NMT) can achieve promising translation quality on resource-rich languages due to end-to-end learning. However, the widely-used NMT system only focuses on modeling the inner mapping from source to target without resorting to external knowledge. In this paper, we take English-Chinese translation as a case study to exploit the use of knowledge graph (KG) in NMT. The main idea is utilizing the entity relations in knowledge graph as constraints to enhance the connections between the source words and their translations. Specifically, we design two kinds of constraints. One is monolingual constraint that employs the entity relations in KG to augment the semantic representation of the source words. The other is bilingual constraint which enforces the entity relations between the source words to be shared by their translations. In this way, external knowledge can participate in the translation process and help to model semantic relationships between source and target words. Experimental results demonstrate that our method outperforms the state-of-the-art system.

Keywords: Neural Machine Translation, Knowledge-constrain, Knowledge Graph.

1 Introduction

With the rapid development of neural machine translation (NMT), we have witnessed the success of various NMT frameworks based on different neural network architectures such as recurrent neural network [2, 13], convolutional neural network [7] and purely attention network [15]. Due to the powerful modeling capacity of these networks, promising translation quality can be achieved in several resource-rich language pairs. However, the conventional methods only focus on how to model the relationship between parallel sentences without resorting to any external knowledge (e.g. knowledge graph, KG). As a result, previous methods lack the ability to figure out the similar relations between (go, went) and (eat, ate). It is also incapable to tell the distinction between (king, man)

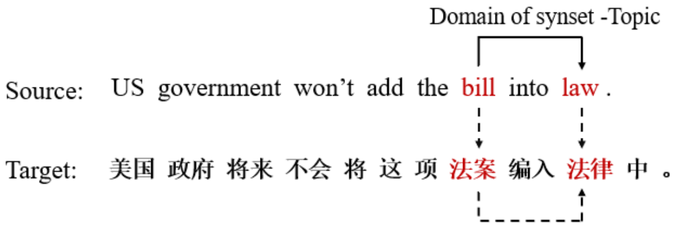
Domain of synset -Topic
Source: US government won't add the bill into law .
Target: 美国 政府 将来 不会 将 这 项 法案 编入 法律 中 。

Fig. 1. An example of English-Chinese translation. The source word "bill" and "law" share the same relationship as the target ones

and (queen, women). Intuitively, the semantic relations between words should be maintained during translation while current methods cannot guarantee this.

To address the above problem, we attempt to take advantage of the rich entity relations embodied in the knowledge graph to guide the translation process. The involving of KG strengthens the semantic relations between words and bridges rare words with common ones. We consider extracting multiple structured information from the existing knowledge graph to connect words with different relations. Since the knowledge in graph is distributed in various domains, including syntactic relations or other common-sense information, we can apply diverse knowledge to NMT.

We first extract from knowledge graph the triplets, consisting of a head word, a tail word and their relation, and then convert them to a computable format. To fully explore the usage of the knowledge graph in machine translation, we design two approaches using entity relations as constraints. One is monolingual constraint, which utilizes the entity relations to influence only the source side. Specifically, the monolingual constraint requires the embedding of the source words to hold the semantic relationship provided by the knowledge graph. The other is bilingual constraint that model relation equivalence between source words and their translations. Specifically, the bilingual constraint enforces the semantic relation between the source words should be exactly maintained by their corresponding translations. Figure 1 illustrates an example of English-Chinese translation. The relation between Chinese words 法案(bill) and 法律(law) in the target language should be the same as that between"bill" and "law" in the source language. Both of the monolingual and bilingual relation constrains are modeled during the training process and make the NMT system much more knowledgeable.

Due to availability of large-scale English knowledge graph, we perform English to Chinese translation task in the experiments to verify the effectiveness of our method. We expect the NMT training process would benefit a lot from the supplementary English KG under monolingual and bilingual constraints respectively. The extensive experimental results demonstrate that our method can outperform the state-of-the-art Transformer model in translation quality.

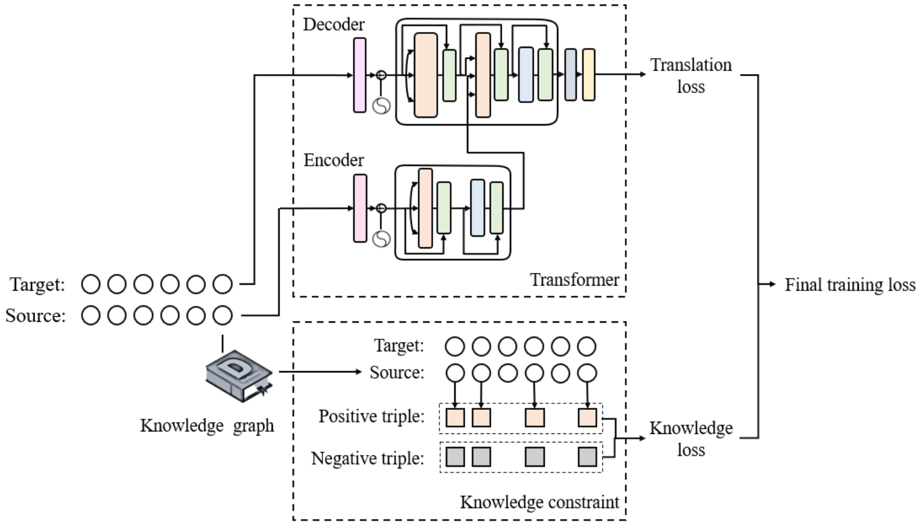

Fig. 2. The framework of our model. The Transformer part and knowledge constraint are independent before calculating final loss.

To further figure out how the external knowledge influences translation, we investigate whether the learned word embeddings really encode the semantic relations imposed by KG. Analogy prediction test is designed and implemented on the word embeddings. In this test, the head word and relation are given, our model selects the proper tail word from the candidate set. Compared to the baseline, our model has the ability of analogy and reasoning to some extent.

The main contributions of this paper are as follows:

- The knowledge graph is first applied in NMT to improve the translation quality.
- We design both monolingual and bilingual constraints to fully exploit the KG entity relations in the training procedure of NMT.
- The experiments on English-Chinese translation task show that both monolingual and bilingual constraints could achieve moderate improvements over the strong Transformer baseline.
- Our method significantly decreases the appearance of unknown words (UN-K). The number of UNKs drops by 30.84% in NIST 2005 and the average reduction is 15% in other two test sets.

2 Neural Machine Translation

As shown in Figure 2, our model poses knowledge constraint on word representations which can be implemented under various NMT architectures. In this paper, we utilize purely-attention transformer architecture which is shown in Figure 3.

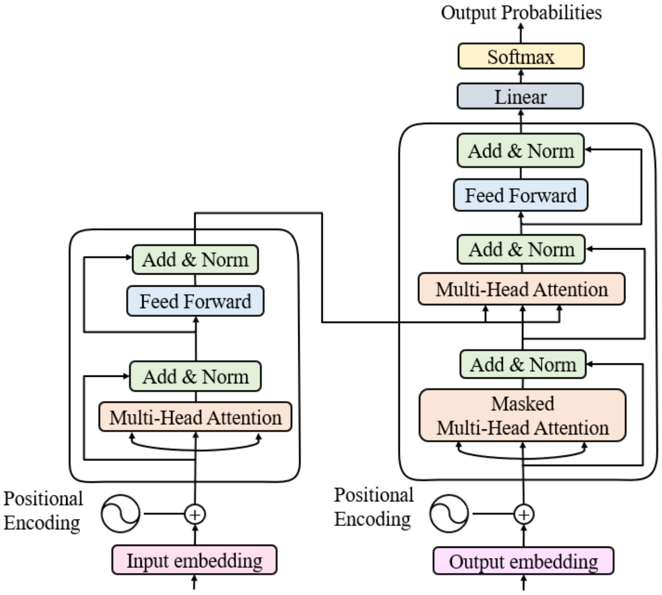

Fig. 3. The architecture of transformer.

Given the source sentence $X = \{x_1, x_2, ...x_n\}$ and the target one $Y = \{y_1, y_2, ...y_m\}$, this model abandons the idea of encoding successively and choose to operate self-attention mechanism over inputs repeatedly to obtain context information. Then, decoder also performs self-attention themselves and implement a multi-head attention upon the output of encoder to generate translations.

The encoder is a stack of six identical layers, each of which includes two sub-layers. A multi-head self-attention layer is set as the first sub-layer and a simple position-wise fully connected feed-forward network is the second one. Besides, a residual connection around each sub-layer is performed and followed by a normalization layer.

The decoder is also composed of six identical layers which have the same sub-layers as those in encoder. In addition, a multi-head attention over the encoder outputs is performed to help produce target translations.

Given the training parallel data $\left\{(X^{(z)}, Y^{(z)})\right\}_{z=1}^{Z}$, the final training loss function are all similarly defined as the conditional log-likelihood in despite of varied framework:

$$L\left(\theta\right) = -\frac{1}{Z}\sum_{z=1}^{Z}\sum_{i=1}^{m} logp\left(y_i^{(z)}|y_{<i}^{(z)}, x^z, \theta\right) \tag{1}$$

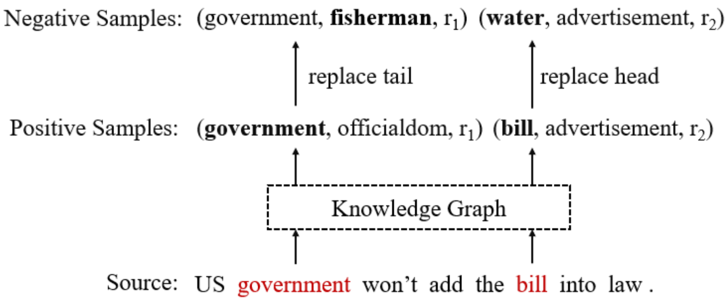
Negative Samples:
(government, **fisherman**, r1)
(**water**, advertisement, r2)
replace tail
replace head
Positive Samples:
(**government**, officialdom, r1)
(**bill**, advertisement, r2)
Knowledge Graph
Source:
US government won't add the bill into law .

Fig. 4. An example of bilingual constraint. The source word "government" has relation r_1 with "officialdom" and "bill" has relation r_2 with "advertisement" in KG. We replace the head word "bill" and the tail one "officialdom" to construct negative samples.

3 Model Description

Similar to other NMT systems, the architecture mentioned above only concentrates on modeling inner mapping between parallel sentences. In this paper, we integrate entity relations, which are transformed to embedding constraint patterns independently, to strengthen the semantic relations between source and target words. In this section, we first introduce how to formalize embedding constraint on entity words. Then, two kinds of constraints, monolingual constraint and bilingual one, are designed to make knowledge assist training procedure.

3.1 Embedding Constraint for Relation Triples

We extract a set of fact triples, T, from knowledge graph. Each triplet is composed of a head word, a tail word and their relation. Referring to Lin et al.[9], we use embedding to denote the elements of triples which are mapped to the corresponding hyper space by mapping matrix as follows. The subtraction of the mapped entity vectors is forced to be close to the mapped relation vector. Specifically, the fact triple with relation r can be formulated as:

$$Tuple_r = (head, tail, r) \tag{2}$$

The entities and relation are mapped to embeddings, e_{head}, e_{tail} and e_r, and the following equation holds for each triple:

$$e_{head} M_r \approx e_{tail} M_r + e_r \tag{3}$$

where M_r is the mapping matrix which is specific to relation r. We score how well the elements of a triplet match each other by:

$$f_r\left(head, tail\right) = \left|e_{head} M_r + e_r - e_{tail} M_r\right| \tag{4}$$

3.2 Monolingual Constraint

In monolingual constraint, we only employ the entity relations to only influence the semantic embedding of the source words. Given the source sentence $X = \{x_1, x_2, ...x_n\}$, we extract the triples whose head word occurs in the source sentence, $S_{POS} = \{(head, tail, r) | h \in x, (h, tail, r) \in T\}$. For each positive triple in S_{pos}, we replace the head or tail word to construct the negative samples as $(head^{'}, tail^{'}, r)$.

As Figure 4 illustrated, "government" has the relation r_1 with "officialdom" which is the out-of-sentence word. $(``government'', ``officialdom'', r_1)$ could be seen as positive sample and $(``government'', ``fisherman'', r_1)$ is constructed to be negative one by replacing tail word "officialdom" with "fisherman".

Instead of random replacement which may introduce false negative labels, we choose to select replacement by probabilities. A different sampling method in Wang et al.[16] is adapted where the head entity is more likely to be swapped if the relation is one-to-many.

The loss function of monolingual constraint can be written as:

$$L(x^{(z)}, y^{(z)}) = \frac{1}{N} \sum_{(head,tail,r) \in S_{pos}} max(0, f_r(head, tail) + \gamma - f_r(head^{'}, tail^{'})) \tag{5}$$

where N denotes the number of positive samples and γ is the hyper parameter.

3.3 Bilingual Constraint

In bilingual constraint, we reckon that the relation between source entities should be maintained by their translations. We first extract all triples, whose head and tail words both appear in source sentence, $S_{src} = \{(head_{src}, tail_{src}, r) | head_{src} \in x, tail_{src} \in x\}$. Then we align the head and tail words of triples in S_{src} to their translations as $S_{tar} = \{(head_{tar}, tail_{tar}, r) | head_{tar} \in y, tail_{tar} \in y\}$. To minimize the gap between source triplet and aligned target one, the loss function can be set as:

$$L(x^{(z)}, y^{(z)}) = -\frac{1}{N} \sum_{(head_{src},tail_{src},r) \in S_{pos}} |f_r(head_{src}, tail_{src}) - f_r(head_{tar}, tail_{tar})| \tag{6}$$

3.4 Adding Constraint in NMT

Monolingual or bilingual constraints are used to improve semantic word embeddings during NMT training. As shown in Figure 2, the overall loss functions mainly includes two parts: one is from conventional translation and the other is from the entity relation loss:

$$Loss = \frac{1}{Z} \sum_{z=1}^{Z} \sum_{i=1}^{m} log\left(y_i^{(z)} | y_{<i}^{(z)}, x^z, \theta\right) + \alpha \frac{1}{Z} \sum_{z=1}^{Z} L(x^{(z)}, y^{(z)}) + \beta ||M_r||^2 \tag{7}$$

where α,β are hyper parameters and $||M_r||^2$ is set as regularization. In practice, the translation loss and the entity relation loss are optimized iteratively.

4 Experiments

4.1 Dataset

We conduct our experiments on the NIST English-Chinese translation task since there are plenty of knowledge graphs in English. Due to absence of multiple test sets for English-Chinese translation, we construct the test sets from the original NIST Chinese-English dataset in which each Chinese source sentence has four English references. For each instance, we regard first English reference as source sentence and the Chinese sentence as single reference. The evaluation metric is BLEU[12] and we select the character-based BLEU-5 which is suitable for Chinese.

Our training data consists of 2.1M sentences pair. Besides, NIST 2002 dataset and NIST 2005, 2006 and 2008 datasets are selected as our development and test sets, respectively.

For knowledge extraction, we filter triples from Wordnet covering 155K entities and 27 relationships. So as to obtain high-quality triples, we discard some low-frequency relations and dual triples where exchange of head and tail words makes no difference. The final fact triple extraction covers 11 relations, which is utilized to match entities in source and target sentences. In monolingual case, the average matching number of one sentence is 7.54 triples. While in bilingual case, the triples where head and tail entities are both lying in source sentence are selected and Fast Align[6] is used to match target side words. However, because of low matching ratio of bilingual situation, we only make our bilingual experiments on 250K pairs of sentences which cover matched entities.

As for analogy prediction test, we extract 1K triples from Wordnet and arbitrarily select another four words for each triple as tail word candidates. We also have to ensure that the given head word, relation and five tail word candidates are contained in dictionaries.

4.2 Training Details

We perform all the experiments using Tensor2tensor, an open-source tool provided by Google. The settings of the training procedure and the hyper parameters are similar to "transformer big single GPU" mode in Tensor2tensor. In detail, we set batch size as 1024, hidden layer as 512 and training step as 300K. We limit the vocabularies to the words whose frequencies are more than 20 and also construct relation vocabulary covering 11 relations. During training, we set hyper parameters α among $\{0.035, 0.06, 0.1, 0.2\}$, β among $\{0.0001, 0.001, 0.002, 0.006\}$ for monolingual constraint. In bilingual constraint, we select α among $\{0.5, 1, 1.5, 2\}$, β among $\{0.035, 0.07, 0.105, 0.15\}$, γ is set to 10.

4.3 Results on English-Chinese Translation

Monolingual We list BLEU-5 scores of our monolingual model in Table 1. Compared with the NMT baseline implemented under transformer architecture with the same setting, our model get an average improvement of +0.947 BLEU over the state-of-art performance.

Table 1. Translation results (BLEU-5 score) for English-Chinese task in monolingual constraint.

System	NIST05	NIST06	NIST08	Ave
NMT Baseline	24.227	24.72	17.67	22.613
Our Model	$24.529^{+0.302}$	$24.91^{+0.19}$	$18.793^{+1.123}$	$23.560^{+0.947}$

Moreover, we find that our method achieves a substantial decrease of the number of <UNK>. As shown in Table 2, the <UNK> frequency drops by 30.84% in NIST05 and other test sets all enjoy a decrease to some extent. The reason behind is that relation modeling provides sufficient training for words which is hard to handle before. As shown in Table 4, “blew” is not familiar to original model which is represented as <UNK>. In our method, “blew” is the past form of “blow” and that relation (blew, blow, past-style of verb) is fully modeled to get better embedding representations.

Table 2. Statistics of the descent rate for UNK in monolingual constraint.

System	NIST05	NIST06	NIST08	Ave
Our Model	30.84%	7.27%	23.15%	19.25%

Bilingual When testing the effect of bilingual constraint, we only filter the specific sentences of NIST05, NIST06, NIST08, which cover at least one trained entity. As shown in Table 3, the enhanced model with bilingual constraint outperforms the baseline system by an average BLEU score of 0.675.

Table 3. Translation results (BLEU-5 score) for English-Chinese task in bilingual case for the subset of NIST 05, 06 and 08.

System	NIST05$'$	NIST06$'$	NIST08$'$	Ave$'$
NMT Baseline	21.519	21.693	13.872	19.594
Our Model	$22.464^{+0.945}$	$22.343^{+0.65}$	$14.105^{+0.233}$	$20.269^{+0.675}$

Table 4. Translation examples, where our method is more capable of obtaining accurate translations than baseline NMT in monolingual constrain.

Source	Palestinian suicide bomber blew up bus, 7 dead and 30 injured.
Reference	巴赶死队成员引爆巴士7人死30人伤。
Baseline	巴勒斯坦自杀炸弹杀手<UNK>公共汽车7死30伤。
Our Model	巴勒斯坦自杀炸弹杀手引爆公共汽车7人死亡30人受伤。
Source	Days of heavy snow in Europe left may dead and transportation disrupted.
Reference	欧洲连日大雪多人死亡交通中断。
Baseline	欧洲大雪<UNK>，许多人死亡，交通中断。
Our Model	欧洲大雪几天，许多人丧生，交通受阻。
Source	The verification department found quite a few fake college diplomas.
Reference	鉴定部门发现，大学生们的证书中也有不少是假货。
Baseline	<UNK>发现不少假大学文凭。
Our Model	核查部门发现了不少假大学文凭。

4.4 Results on Analogy Prediction Test

To further investigate whether the learned word embeddings actually encode the semantic relations imposed by KG, we implement analogy prediction test on the trained embeddings, which aims to predict the missing head or tail word for a relation fact triple (head, tail, r), mentioned in Mikolov et al.[10].

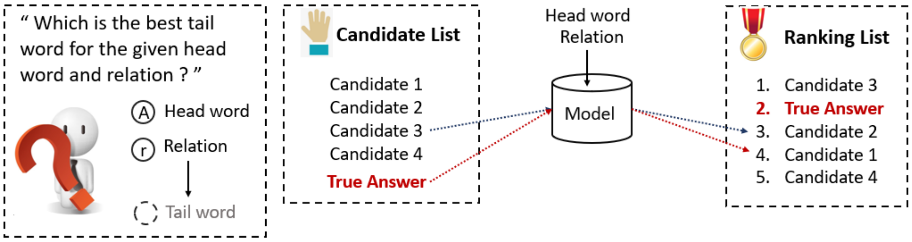

Fig. 5. An overview of analogy prediction task.

As illustrated in Figure 5, given the head word and relation, our model has to select the proper tail word from five candidates by calculating the score as:

$$score(x) = |e_{head}M_r - e_xM_r + e_r| \tag{8}$$

We select 1K triples and swap their tail words. Our model ranks the candidates by their scores to evaluate the probabilities that being the true tail word. From the result shown in Table 5, the gold answers ranking the first account for 28.3% while the random baseline is 20%. It indicates that our model has some capacity of analogy prediction compared to randomly selection.

Table 5. The ranking results for gold answers in analogy prediction test.

Rank	1	2	3	4	5
Number	283	169	148	178	222

5 Related Work

Recently, neural machine translation has attracted more and more attention. A novel pure attention architecture, Transformer, is proposed by Vaswani et al.[15], which achieves state-of-art performance and has much faster training speed. Our model is also implemented under the Transformer framework.

Previous work mainly put emphasis on modeling the mapping function from source sentence to target sentence and external knowledge (e.g. knowledge graph) is in the absence of translate process. However, there are still some researches on how to enable knowledge to benefit translation. Li et al.[8] employed the "synonym" and "hypernym" relations extracted from Wordnet to find proper replacements for low-frequency words. Zou et al.[22] proposed that the embeddings of bilingual aligned words should be closer. Semantic gap between the source language string and its translation is minimized [19, 20]. Synonyms extracted from dictionaries have been adopted to transform low-frequency word to adequate sequences [21]. Compared to their methods, we consider more diverse relation types between entities.

With respect to knowledge use in NMT, knowledge graph is a good choice to extract structured information. Knowledge graph is created to model the entities and their relations in the real world. It is widely used in question answering system and recommendation systems. To date, the main knowledge graphs are Freebase[3] , Wordnet, Google Knowledge Vault[5] and DBPedia[1], which are mainly developed in English. In the graph, entities are linked by different relationships so that no one would be isolated from others. To better model the elements in the graph and address the link prediction task, Bordes et al.[4] presented a method named TransE which represents entity and relation with embedding and defines the arithmetic relations between the embeddings of head word, tail word and relation. Lin et al.[9] pointed out that entity and relation embeddings should be built in separate entity space and relation spaces. Besides , there existed many other model methods on entity and relation[17, 18, 16, 14, 11].

6 Conclusion

In this paper, we propose a method that integrates the external knowledge into NMT, aiming to augment the connections between the source words and their translations. We design two methods using entity relations as constraints to fully explore the usage of the knowledge graph in NMT. The first one is monolingual constraint which requires the embedding of the source words to hold the semantic relationship provided by KG. The second is bilingual constraint that enforces the semantic relation between the source words to be exactly maintained by their corresponding translations. Experimental results demonstrate that our method obtains much improvements over the strong NMT baseline.

Acknowledgements

The research work described in this paper has been supported by the National Key Research and Development Program of China under Grant No. 2016QY02D0303 and the Natural Science Foundation of China under Grant No. 61673380.

References

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: The Semantic Web, International Semantic Web Conference, Asian Semantic Web Conference, ISWC 2007 + Aswc 2007, Busan, Korea, November. pp. 722–735 (2007)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. International Conference on Learning Representations (2015)
3. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase:a collaboratively created graph database for structuring human knowledge. In: SIGMOD Conference. pp. 1247–1250 (2008)
4. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: International Conference on Neural Information Processing Systems. pp. 2787–2795 (2013)
5. Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., Zhang, W.: Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 601–610 (2014)
6. Dyer, C., Chahuneau, V., Smith, N.A.: A simple, fast, and effective reparameterization of ibm model 2. Proc Naacl (2013)
7. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning. arXiv:1705.03122 (2017)
8. Li, S., Xu, J., Miao, G., Zhang, Y., Chen, Y.: A semantic concept based unknown words processing method in neural machine translation. Natural Language Processing and Chinese Computing (2017)
9. Lin, Y., Liu, Z., Zhu, X., Zhu, X., Zhu, X.: Learning entity and relation embeddings for knowledge graph completion. In: Twenty-Ninth AAAI Conference on Artificial Intelligence. pp. 2181–2187 (2015)
10. Mikolov, T., Yih, W.T., Zweig, G.: Linguistic regularities in continuous space word representations. In HLT-NAACL (2013)
11. Nickel, M., Rosasco, L., Poggio, T.: Holographic embeddings of knowledge graphs. National Conference on Artificial Intelligence pp. 1955–1961 (2016)
12. Papineni, Kishore, Roukos, Salim, Ward, Todd, Zhu, WeiJing: Bleu: a method for automatic evaluation of machine translation. Meeting of the Association for Computational Linguistics **4**(4), 307–318 (2001)
13. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: International Conference on Neural Information Processing Systems. pp. 3104–3112 (2014)

14. Trouillon, T., Welbl, J., Riedel, S., Gaussier, E., Bouchard, G.: Complex embeddings for simple link prediction. International Conference on Machine Learning pp. 2071–2080 (2016)
15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv:1706.03762v5 (2017)
16. Wang, Z., Zhang, J., Feng, J., Chen, Z.: Knowledge graph embedding by translating on hyperplanes. In: Twenty-Eighth AAAI Conference on Artificial Intelligence. pp. 1112–1119 (2014)
17. Xiao, H., Huang, M., Hao, Y., Zhu, X.: Transa: An adaptive approach for knowledge graph embedding. Computer Science (2015)
18. Xiao, H., Huang, M., Zhu, X.: Transg : A generative model for knowledge graph embedding. In: Meeting of the Association for Computational Linguistics. pp. 2316–2325 (2016)
19. Zhang, J., Liu, S., Li, M., Zhou, M., Zong, C.: Bilingually-constrained phrase embeddings for machine translation. Meeting of the Association for Computational Linguistics **1**, 111–121 (2014)
20. Zhang, J., Liu, S., Li, M., Zhou, M., Zong, C.: Mind the gap: machine translation by minimizing the semantic gap in embedding space. national conference on artificial intelligence pp. 1657–1663 (2014)
21. Zhang, J., Zong, C.: Bridging neural machine translation and bilingual dictionaries. arXiv: Computation and Language (2016)
22. Zou, W.Y., Socher, R., Cer, D.M., Manning, C.D.: Bilingual word embeddings for phrase-based machine translation. Empirical Methods in Natural Language Processing pp. 1393–1398 (2013)