entropy

Article

Improving Neural Machine Translation by Filtering Synthetic Parallel Data

Guanghao Xu [1], Youngjoong Ko [2,*] and Jungyun Seo [1]

[1] Department of Engineering, Computer Science, Sogang University, Seoul 04107, Korea; guanghao412@gmail.com (G.X.); seojy@sogang.ac.kr (J.S.)
[2] Applied Data Science, Sungkyunkwan University, Suwon 16419, Korea
* Correspondence: yjko@skku.edu or youngjoong.ko@gmail.com

Received: 23 October 2019; Accepted: 10 December 2019; Published: 11 December 2019

Abstract: Synthetic data has been shown to be effective in training state-of-the-art neural machine translation (NMT) systems. Because the synthetic data is often generated by back-translating monolingual data from the target language into the source language, it potentially contains a lot of noise—weakly paired sentences or translation errors. In this paper, we propose a novel approach to filter this noise from synthetic data. For each sentence pair of the synthetic data, we compute a semantic similarity score using bilingual word embeddings. By selecting sentence pairs according to these scores, we obtain better synthetic parallel data. Experimental results on the IWSLT 2017 Korean→English translation task show that despite using much less data, our method outperforms the baseline NMT system with back-translation by up to 0.72 and 0.62 BLEU points for tst2016 and tst2017, respectively.

Keywords: neural machine translation; back translation; bilingual word embeddings; synthetic data filtering

1. Introduction

Recent advances in neural machine translation (NMT) have achieved human parity on several language pairs given large-scale parallel corpora [1,2]. However, for many language pairs, the amount of parallel corpora is limited; this is a major challenge in building high-performance machine translation (MT) systems [3]. By contrast, there are plenty of monolingual data, which are easier to obtain.

Many approaches have been proposed to improve MT systems by leveraging monolingual data [4,5]. Sennrich et al. [6] proposed a back-translation approach to expand a parallel training corpus with synthetic parallel data. In this approach, the synthetic parallel data are constructed by translating the target-language monolingual data into the source language with a backward translation (target-to-source) model trained by a given parallel training corpus. Although this approach can generate a large amount of synthetic parallel data, there is no guarantee of its quality.

Regarding synthetic data filtering, Imankulova et al. [7] attempted to filter out those low-quality sentence pairs from the synthetic parallel data. To measure the quality of synthetic sentence pairs, they first translated synthetic source sentences to construct synthetic target sentences by using a forward translation (source-to-target) model. Then, for each sentence pair, the sentence-level BLEU [8] score between the target-language monolingual sentence and the target-language synthetic sentence was calculated. Finally, sentence pairs of the lower score were filtered out from the synthetic parallel corpus. By filtering out noisy sentence pairs, they obtained improvements over the baselines on several low-resourced translation tasks. However, they observed that translation performance did not improve when the size of monolingual data was large, i.e., over 1 million sentences. Furthermore, to calculate the sentence-level BLEU scores, they built an additional translation model to generate the target-language synthetic sentences.

Following the shared task on parallel corpus filtering introduced by Koehn et al. [9] at WMT2018 (Third Conference on Machine Translation), in this paper, we propose a simple and effective approach to filter out noisy sentence pairs from synthetic parallel data. Our approach is based on sentence-level cosine similarities of two sentence vectors, i.e., vector representations of the synthetic source sentence and the monolingual target sentence. We calculate the sentence vectors by averaging the word embeddings of each sentence. In addition, to locate the sentence vectors in a common vector space, we learn bilingual linear mappings between word embeddings of the source and the target language. The proposed method has two advantages: (1) no additional translation models are required to generate synthetic target sentences, and (2) semantic information of words in both synthetic and monolingual sentences are considered by using both source and target word embeddings. To the best of our knowledge, no previous works have investigated the similarity of the synthetic source and the target sentence in the context of synthetic parallel corpus filtering.

The remainder of this paper is structured as follows. In Section 2, we describe related research. In Section 3, we introduce our proposed filtering method. In Section 4, we present the experimental setup. In Section 5, we discuss the results of our experiment. Finally, in Section 6, we conclude the paper and suggest future work.

2. Related Work

Most of the methods of learning bilingual embeddings are supervised and rely on a small bilingual dictionary of a few thousand word pairs. Mikolov et al. [10] first proposed a cross-lingual embedding mapping method, which maps word embeddings in two languages by learning a linear transform. Xing et al. [11] found inconsistencies in the objective function of the linear transform, and proposed to constrain the linear transform as an orthogonal transform. Luong et al. [12] proposed a joint model that used both the context co-occurrence information through the monolingual component and the meaning equivalent signals from the bilingual constraint. They showed that the model was capable of learning bilingual representations of two languages, simultaneously preserving the monolingual clustering structures in each language. Artetxe et al. [13] proposed a framework of learning bilingual mappings of word embeddings, which generalized previous research.

Several studies examined the context of learning bilingual embeddings in a semi-supervised or unsupervised scenario, where the bilingual dictionary was much smaller. Artetxe et al. [14] proposed a self-learning approach that induced a new bilingual dictionary iteratively, achieving comparable results with only 25 word pairs. Conneau et al. [15] showed that they could build a high-quality bilingual dictionary without cross-lingual supervision. Their method leveraged both the domain-adversarial training approach and an iterative refinement procedure. Artetxe et al. [16] proposed a new unsupervised approach to learn cross-lingual embedding mappings by exploiting the structural similarity of the embeddings.

There are several studies on handling noise in parallel data. For example, Taghipour et al. [17] used a probability density estimation algorithm to detect outliers in parallel data. Cui et al. [18] proposed a graph-based random walk algorithm to compute the quality score of each sentence pair. Junczys-Dowmunt [19] introduced a dual conditional cross-entropy filtering, which computes cross-entropy scores based on the two translation models trained on clean data. These studies focused on filtering noise in the parallel data crawled from the web, instead of synthetic parallel data.

3. Proposed Method

3.1. Neural Machine Translation

A standard state-of-the-art NMT system follows the encoder-decoder framework. It includes two main components: an encoder network and a decoder network [20]. Given a source and target sentence pair (X, Y), where $X = x_1, ..., x_M$ and $Y = y_1, ..., y_N$, the encoder network first takes source sentence X as an input and generates a list of fixed-size vectors $S = s_1, ..., s_M$, whose size is the length of the source sentence. Next, the decoder network predicts each token sequentially by maximizing the conditional probability:

$$p\left(y_t | y_1, ..., y_{t-1}, S\right) = \text{softmax}\left(\mathbf{W}_o\, \mathbf{h}_t\right),$$

where $\mathbf{W}_o$ is the weight of the output softmax layer and $\mathbf{h}_t$ is the target hidden state at time step t.

Given a parallel corpus D, the training objective is to minimize the cross-entropy loss:

$$\mathcal{L} = \sum_{(X,Y) \in D} \sum_{t=1}^{N} \log p\left(y_t | y_1, ..., y_{t-1}, S\right).$$

3.2. Back-Translation for NMT

Back-translation is a technique that employs target-language monolingual data in training the NMT system without changing its network architecture. Given a sentence-aligned parallel dataset $D_p = \left\{\left(X_p^n, Y_p^n\right)\right\}_{n=1}^{N}$ and a monolingual dataset in the target language $D_{tm} = \left\{Y_{tm}^{(m)}\right\}_{m=1}^{M}$, the process of back-translation includes the following steps. First, a translation model in the reverse direction $\text{NMT}_{Y \to X}$ is trained with the parallel dataset D_p. Second, with the translation model $\text{NMT}_{Y \to X}$, the target-language monolingual dataset D_{tm} is back-translated into the source-language translations $D_{st} = \left\{X_{st}^{(m)}\right\}_{m=1}^{M}$, which is then paired with D_{tm}, making up a synthetic parallel dataset $D_{syn} = \{(X_{st}^m, Y_{tm}^m)\}_{m=1}^{M}$. Third, synthetic parallel dataset D_{syn} and real parallel dataset D_p are combined to train the main translation model $\text{NMT}_{X \to Y}$.

3.3. Synthetic Parallel Data Filtering with Bilingual Word Embeddings

The filtering method introduced in this section is our main contribution. Our filtering method relies on cosine similarities of sentence embedding vectors in a common vector space. For each sentence x, we create its sentence embedding vector by accumulating word vectors w_1 to $w_{|x|}$, which are then averaged to form a single mean vector representation.

$$s_x = \frac{1}{|x|} \sum_{t=1}^{|x|} w_t$$

For each sentence pair (x, y) in the synthetic parallel corpus, cosine similarity of s_x and s_y is computed as

$$\text{similarity}\left(s_x, s_y\right) = \frac{s_x \cdot s_y}{|s_x||s_y|}.$$

Because the two sentences in each pair are written in different languages, it is necessary to ensure that the vector representations of these sentences are located in the same vector space.

A common approach to solve this problem is by using bilingual word embeddings. Following the work in [13,14,16], we first train word embeddings X and Z for the source and target language, respectively. Then, with a small bilingual dictionary, we learn a linear mapping W that minimizes the sum of squared Euclidean distances:

$$\underset{W}{\text{argmin}} \sum_{i=1}^{n} \left\| X_i W - Z_i \right\|^2,$$

where X_i and Z_i are the vector representations of word pairs in the bilingual dictionary.

Once the similarity scores of all sentence pairs are computed, we use a threshold value t to eliminate the sentence pairs with the scores below the threshold. The threshold value is computed by linearly scaling the similarity scores into the range of $[0, 1]$.

4. Experimental Setup

4.1. Datasets and Data Preprocessing

For Korean→English experiments, we used parallel training data released in IWSLT2017 [21] (the translation of TED talks). Besides, we used tst2016 and tst2017 as evaluation datasets (Available online: https://wit3.fbk.eu/). Monolingual data (English) for back-translation were obtained from the WMT2016 German–English news translation task. Dataset statistics is shown in Table 1.

Table 1. Statistics for the Korean–English translation datasets.

Dataset	Source	Target	Sentences
IWSLT2017	Korean	English	207 K
WMT2016	-	English	4.5 M
tst2016	Korean	English	1024
tst2017	Korean	English	1036

The English sentences were tokenized and true-cased with Moses [22] preprocessing scripts. The Korean sentences were tokenized with Komoran (Available online: http://konlpy.org/en/) [23] tokenizer. We removed sentence pairs longer than 50 words and learned a joint source and target byte-pair encoding [24] with 32,000 merge operations.

All translation results reported in this paper were calculated in terms of single reference case-insensitive BLEU measured with Moses' `multi-bleu.perl` script (Available online: https://github.com/moses-smt/mosesdecoder).

4.2. Models and Hyperparameters

The NMT system we used for evaluation is the OpenNMT [25] implementation of the Transformer [26] model. We followed the settings of the base model described in the paper, i.e., 6 attention blocks in the encoder and decoder, the embedding of size 512, and feed-forward dimension 2048. We used 8 attention heads, and we averaged the last 10 checkpoints, which were saved every 10,000 training steps.

The NMT system used for back-translation was an encoder–decoder model based on a 4-layer recurrent neural network (RNN). Specifically, we used the long short-term memory (LSTM) [27] and the attention mechanism proposed by Luong et al. [28]. We set hidden units to 1024, dropout rate to 0.2, and mini-batch size to 128. We trained the model with the stochastic gradient descent algorithm using a learning rate of 1.0, and we generally followed the learning rate decay scheme stated in [1].

The bilingual word embedding model used in our filtering method was obtained as follows. First, we trained word embeddings for Korean and English with fastText toolkit (Available online: https://fasttext.cc/) [29] on Wikipedia data (Available online: https://dumps.wikimedia.org/). Next, we created a list of English words by selecting the top 4500 most frequent words in the English Wikipedia data; function words and stop words were not included in the list. Subsequently, a bilingual (Korean and English) speaker translated all English words into Korean. Finally, we used existing approaches (Available online: https://github.com/artetxem/vecmap) to learn linear transformation matrix W with the word embeddings and the bilingual dictionary.

5. Experimental Results and Discussion

5.1. Quality of Bilingual Word Embeddings

To evaluate the quality of bilingual word embeddings, we created a word translation task that considered the translation accuracy of the given source words. The test set used in this task contains 500 word pairs that were uniformly selected from the bilingual dictionary. The bilingual word embeddings were obtained by applying existing approaches: Supervised [13], Identical [14], and Unsupervised [16]. These approaches mainly differ in which bilingual word pairs are used in learning linear transformation.

Specifically, the Supervised method learns a mapping using all word pairs in a bilingual dictionary, the Identical method uses identical character strings as bilingual signal, and the Unsupervised method exploits the structural similarity of the embeddings instead of a bilingual dictionary.

Table 2 shows the quality of bilingual word embeddings in terms of word translation accuracy. As shown in Table 2, the supervised mapping method, trained with a bilingual dictionary of 4000 word pairs, achieved 42.60% accuracy, outperforming the other two approaches in our experiment. Therefore, we decided to choose the supervised method to build bilingual word embeddings in the following experiments.

Table 2. Word translation accuracy of bilingual word embeddings on Korean→English word translation task.

Bilingual Mappings	**Accuracy**
Supervised	42.60%
Identical	41.58%
Unsupervised	40.16%

5.2. Size of Synthetic Datasets

Sennrich [6] showed that the translation performance decreases if the size of synthetic data is too large compared to real data. Moreover, Fadaee and Monz [30] found that the model trained on 1:4 real-to-synthetic ratio of training data achieved slight improvements over the model trained on 1:1 training data. Because the size of real parallel data used in our experiments is relatively small, we explored various ratios of synthetic data to test which ratio achieves the best results.

Table 3 presents the translation performance of the systems trained on different ratios of the training data. The baseline model was trained on only real parallel data, whereas the "+ synthetic" models were trained on concatenated real and synthetic data. All models trained with additional synthetic data significantly outperformed the baseline model. In addition, models trained with synthetic data of ratio 1:5 outperformed the ratio 1:1 by a large margin. It is in line with the findings of Fadaee and Monz [30]. To our surprise, the 1:10 ratio of real-to-synthetic data performed best in our experiments. Hence, when the size of the real parallel corpus is relatively small, more synthetic data is required to obtain the best translation performance.

Table 3. Korean→English translation performance (BLEU) on IWSLT test sets (TED talks) with different ratios of *real:syn* data.

Model	**Size**	**tst2016**	**tst2017**
Baseline	207 K	14.46	12.84
+ synthetic (1:1)	414 K	15.55	13.67
+ synthetic (1:5)	1.2 M	17.44	15.24
+ synthetic (1:10)	2.2 M	**18.01**	**15.48**
+ synthetic (1:20)	4.2 M	16.26	14.03

5.3. Quality of Filtered Synthetic Data

Subsequently, we analyze the quality of synthetic data filtered on two different approaches: "Sent-BLEU" and "Sent-BiEMB." For this experiment, we sorted all the sentence pairs in the filtered synthetic data by their similarity scores. Next, we selected the top-ranked 200,000 and 400,000 sentence pairs and constructed new datasets: Top200k and Top400k. Afterward, we trained two NMT systems for each dataset and evaluated their performances on the test sets. The "Sent-BLEU" filtering method proposed by [7] removed noisy synthetic data based on sentence-level BLEU scores. The scores were calculated using the monolingual target sentences as a reference and synthetic target sentences as candidates. The synthetic target sentences were generated by translating source sentences in the synthetic parallel data into the target language. Here, the "Sent-BiEMB" is our proposed filtering method described in Section 3.3. In this experiment, the real parallel data were excluded.

As shown in Table 4, for Top200k synthetic data, the "Sent-BiEMB" model achieves 8.34 and 7.33 BLEU points, outperforming the "Sent-BLEU" by +2.64 and +2.04 BLEU points, on tst2016 and tst2017. Similar improvements are observed for Top400k synthetic data. The result indicates that our proposed method "Sent-BiEMB" is more effective than "Sent-BLEU" for filtering noise in synthetic data.

Table 4. Quality of filtered synthetic data in terms of translation performance of (BLEU) on IWSLT test set. Systems are trained using only synthetic parallel data filtered with Sent-BLEU and Sent-BiEMB.

Model	Synthetic Data	tst2016	tst2017
Sent-BLEU	Top200k	5.70	5.27
Sent-BiEMB	Top200k	**8.34 (+2.64)**	**7.33 (+2.06)**
Sent-BLEU	Top400k	8.41	7.38
Sent-BiEMB	Top400k	**10.05 (+1.64)**	**9.03 (+1.65)**

5.4. Performance of Proposed Method with a Combination of Real and Synthetic Data

In this section, we investigate the effects of different filtering methods on translation performance. The results are shown in Table 5. All models were trained on a concatenated real parallel data with filtered synthetic parallel data. The baseline was the best model trained on 1:10 real-to-synthetic ratio of training data described in Section 5.2.

Table 5. Korean→English translation performance of (BLEU) on IWSLT test sets (TED talks). Systems differ in how the synthetic parallel data is filtered.

Model	Threshold	tst2016	tst2017
Baseline	None	18.01	15.48
Sent-BLEU	0.1	17.97	15.66
Sent-BLEU	0.2	17.67	15.26
Sent-BLEU	0.3	17.45	15.39
Sent-BLEU	0.4	16.93	14.65
Sent-BiEMB	0.1	18.03	15.73
Sent-BiEMB	0.2	18.11	15.70
Sent-BiEMB	0.3	**18.73 (+0.72)**	**16.10 (+0.62)**
Sent-BiEMB	0.4	18.20	15.97

As shown in Table 5, the filtering method based on sentence-level BLEU scores did not improve translation performance. This indicates that sentence-level BLEU is not as reliable as a filtering metric when the size of synthetic data is large. It is also in line with the result in [7].

Meanwhile, all "Sent-BiEMB" models outperformed the strong baseline model on both tst2016 and tst2017. The model with a similarity threshold of 0.3 achieved the best result, outperforming the baseline by +0.72 and +0.62 BLEU points (We have performed a test of significance on improvements of the proposed model over the baseline. The test statics (z-score) of tst2016 and tst2017 are 12.63 and 14.05, respectively. The P-value of both test sets is less than 0.0001. Therefore, we conclude that the gains over the baseline are statistically significant). It confirms that filtering noisy sentence pairs from synthetic parallel data with bilingual word embeddings improves the translation models.

6. Conclusions

In this paper, we proposed a simple approach to filtering noisy sentence pairs from a synthetic parallel corpus generated with back-translation. We measured the sentence-level similarities between the synthetic source and the monolingual target sentence by using bilingual word embeddings.

The distributed representation of words was also considered in the proposed method. We observed gains in translation performance by removing noisy sentence pairs with the proposed method.

In future research, we plan to further analyze the types of noise in the synthetic parallel data generated by back-translation and investigate their effects on translation performance. Additionally, we will evaluate our filtering method on other language pairs.

Author Contributions: Funding acquisition, J.S.; investigation, G.X.; methodology, G.X. and Y.K.; project administration, Y.K.; software, G.X.; supervision, J.S.; writing-original draft, G.X.; writing-review and editing, Y.K.

Funding: This research was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (2017-0-00255, Autonomous digital companion framework and application).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv* **2016**, arXiv:1609.08144. Available online: https://arxiv.org/abs/1609.08144 (accessed on 9 December 2019).
2. Hassan, H.; Aue, A.; Chen, C.; Chowdhary, V.; Clark, J.; Federmann, C.; Huang, X.; Junczys-Dowmunt, M.; Lewis, W.; Li, M.; et al. Achieving Human Parity on Automatic Chinese to English News Translation. *arXiv* **2018**, arXiv:1803.05567. Available online: https://arxiv.org/abs/1803.05567 (accessed on 9 December 2019).
3. Koehn, P.; Knowles, R. Six Challenges for Neural Machine Translation. In Proceedings of the First Workshop on Neural Machine Translation, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 28–39. doi:10.18653/v1/W17-3204. [CrossRef]
4. Lambert, P.; Schwenk, H.; Servan, C.; Abdul-Rauf, S. Investigations on Translation Model Adaptation Using Monolingual Data. In Proceedings of the Sixth Workshop on Statistical Machine Translation, Edinburgh, Scotland, 30–31 July 2011; pp. 284–293.
5. Gulcehre, C.; Firat, O.; Xu, K.; Cho, K.; Bengio, Y. On integrating a language model into neural machine translation. *Comput. Speech Lang.* **2017**, *45*, 137–148. doi:10.1016/j.csl.2017.01.014. [CrossRef]
6. Sennrich, R.; Haddow, B.; Birch, A. Improving Neural Machine Translation Models with Monolingual Data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; Volume 1, pp. 86–96. doi:10.18653/v1/P16-1009. [CrossRef]
7. Imankulova, A.; Sato, T.; Komachi, M. Improving Low-Resource Neural Machine Translation with Filtered Pseudo-Parallel Corpus. In Proceedings of the 4th Workshop on Asian Translation (WAT2017), Taipei, Taiwan, 27 November–1 December 2017; pp. 70–78.
8. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; Association for Computational Linguistics: Straudsburg, PA, USA, 2002; pp. 311–318.
9. Koehn, P.; Khayrallah, H.; Heafield, K.; Forcada, M.L. Findings of the WMT 2018 Shared Task on Parallel Corpus Filtering. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, Belgium, Brussels, 31 October–1 November 2018; pp. 726–739.
10. Mikolov, T.; Le, Q.V.; Sutskever, I. Exploiting similarities among languages for machine translation. *arXiv* **2013**, arXiv:1309.4168. Available online: https://arxiv.org/abs/1309.4168 (accessed on 9 November 2019).
11. Xing, C.; Wang, D.; Liu, C.; Lin, Y. Normalized word embedding and orthogonal transform for bilingual word translation. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, CO, USA, 31 May–5 June 2015; pp. 1006–1011.

12. Luong, T.; Pham, H.; Manning, C.D. Bilingual word representations with monolingual quality in mind. In Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, Denver, CO, USA, 5 June 2015; pp. 151–159.
13. Artetxe, M.; Labaka, G.; Agirre, E. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 5012–5019.
14. Artetxe, M.; Labaka, G.; Agirre, E. Learning bilingual word embeddings with (almost) no bilingual data. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; Volume 1, pp. 451–462.
15. Conneau, A.; Lample, G.; Ranzato, M.; Denoyer, L.; Jégou, H. Word translation without parallel data. *arXiv* **2017**, arXiv:1710.04087. Available online: https://arxiv.org/abs/1710.04087 (accessed on 9 December 2019).
16. Artetxe, M.; Labaka, G.; Agirre, E. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. *arXiv* **2018**, arXiv:1805.06297. Available online: https://arxiv.org/abs/1805.06297 (accessed on 9 December 2019).
17. Taghipour, K.; Khadivi, S.; Xu, J. Parallel corpus refinement as an outlier detection algorithm. In Proceedings of the 13th Machine Translation Summit (MT Summit XIII), Xiamen, China, 19–23 September 2011; pp. 414–421.
18. Cui, L.; Zhang, D.; Liu, S.; Li, M.; Zhou, M. Bilingual data cleaning for smt using graph-based random walk. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, 4–9 August 2013; Volume 2, pp. 340–345.
19. Junczys-Dowmunt, M. Dual conditional cross-entropy filtering of noisy parallel corpora. *arXiv* **2018**, arXiv:1809.00197. Available online: https://arxiv.org/abs/1809.00197 (accessed on 9 December 2019).
20. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems 27*; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2014; pp. 3104–3112.
21. Cettolo, M.; Federico, M.; Bentivogli, L.; Niehues, J.; Stüker, S.; Sudoh, K.; Yoshino, K.; Federmann, C. Overview of the IWSLT 2017 Evaluation Campaign. In Proceedings of the 14th International Workshop on Spoken Language Translation (IWSLT), Tokyo, Japan, 14–15 December 2018; pp. 2–12.
22. Koehn, P.; Hoang, H.; Birch, A.; Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; Zens, R.; et al. Moses: Open Source Toolkit for Statistical Machine Translation. In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (ACL '07), Stroudsburg, PA, USA, 25–27 June 2007; pp. 177–180.
23. Park, E.L.; Cho, S. KoNLPy: Korean natural language processing in Python. In Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology, Chuncheon, Korea, 10–11 October 2014.
24. Sennrich, R.; Haddow, B.; Birch, A. Neural Machine Translation of Rare Words with Subword Units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; Volume 1, pp. 1715–1725. doi:10.18653/v1/P16-1162. [CrossRef]
25. Klein, G.; Kim, Y.; Deng, Y.; Senellart, J.; Rush, A.M. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *arXiv* **2017**, arXiv:1701.02810. Available online: https://arxiv.org/abs/1701.02810 (accessed on 9 December 2019).
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 5998–6008.
27. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. doi:10.1162/neco.1997.9.8.1735. [CrossRef] [PubMed]
28. Luong, T.; Pham, H.; Manning, C.D. Effective Approaches to Attention-based Neural Machine Translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1412–1421. doi:10.18653/v1/D15-1166. [CrossRef]

29. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching Word Vectors with Subword Information. *TACL* **2017**, *5*, 135–146. [CrossRef]
30. Fadaee, M.; Monz, C. Back-Translation Sampling by Targeting Difficult Words in Neural Machine Translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 436–446.