Analysis on Unsupervised Acquisition Process of Bilingual Vocabulary through Iterative Back-Translation

Takuma Tanigawa, Tomoyosi Akiba, Hajime Tsukada
Toyohashi University of Technology
{tanigawa.takuma.fu, akiba.tomoyoshi.tk, tsukada.hajime.hl}@tut.jp

Abstract

In this paper, we investigate how new bilingual vocabulary is acquired through Iterative Back-Translation (IBT), which is known as a data augmentation method for machine translation from monolingual data of both source and target languages. To reveal the acquisition process, we first identify the word translation pairs in test data that do not exist in a bilingual data but do only in two monolingual data, then observe how many pairs are successfully translated by the translation model trained through IBT. We experimented on it with domain adaptation settings on two language pairs. Our experimental evaluation showed that more than 60% of the new bilingual vocabulary is successfully acquired through IBT along with the improvement in the translation quality in terms of BLEU. It also revealed that new bilingual vocabulary was gradually acquired by repeating IBT iterations. From the results, we present our hypothesis on the process of new bilingual vocabulary acquisition where the context of the words plays a critical role in the success of the acquisition.

Keywords: Machine Translation, Bilingual Vocabulary Acquisition, Iterative Back-Translation

1. Introduction

Back-Translation (BT) (Sennrich et al., 2016a) is a common data augmentation method emplyed in Neural Machine Translation. It uses the target-side monolingual data to create a pseudo-bilingual data by employing a reverse-directional translation model. Iterative Back-Translation (IBT) (Hoang et al., 2018; Zhang et al., 2018) is a bi-directional extension of BT. IBT uses two monolingual data of both languages and repeats two processes of creating a pseudo-bilingual data of both directions and updating the translation models of both directions. While previous researches has experimentally demonstrated the effectiveness of IBT, its reason has not been sufficiently elucidated.

In this work, we show that IBT has the unique property of acquiring new bilingual vocabulary solely from two monolingual data and it could be one of the main reasons of its effectiveness. To unveil the acquisition process of IBT, we examine domain adaptation scenarios, where the target domain has many novel words that are never used in the source domain. There's no way initial translation models trained solely from the source domain data can translate them into their correpoonding target words. By applying IBT, we will see if the updated models can translate them correctly.

For each language, we identify the set of the target domain specific words (TDSWs) that do not exist in the corresponding side of the source bilingual data but only in the target monolingual data. From the word aligned sentences in the target domain test data, we search for the TDSW pairs in both languages and define them as Acquirable Bilingual Vocabulary (ABV). By checking if the source TDSWs are correctly translated into the corresponding TDSWs in the target language, we calculate the success rate of word translation with regard to ABV, which is used as our new evaluation metric referred to as Acquisition Rate of Aquirable Bilingual Vocabulary (ARABV). Hu et al. (2019) used a similar metric to see the translation accuracy of unseen in-domain words.

Our experiment was conducted in two different domain adaptation settings on two language pairs. The results indicated that IBT successfully improved the ARABV along with the BLEU scores (Papineni et al., 2002). It also showed that ABV was gradually acquired by repeating the IBT iterations. From the results, we also present our hypothesis on the process of ABV acquisition where the context of the words plays a critical role in the success of the acquisition.

The rest of the paper is organized as follows. Section 2 explains IBT. Section 3 explains procedure for deriving ARABV. Section 4 - 5 describe the experimental setting and results, and detailed analysis on them. Finally, Section 6 presents our hypothesis on the process of ABV acquisition.

2. Related Work

Guo et al. (2021) investigated the compositional generalization ability of IBT on artificial seq2seq tasks. Although the compositional generalization and the acquisition ability of new words can have some relation, it is not so obvious. We directly analyzed the IBT's acquisition ability of new words on machine translation tasks.

Fadaee and Monz (2018) revealed difficult-to-

LREC-COLING 2024, pages 887–892

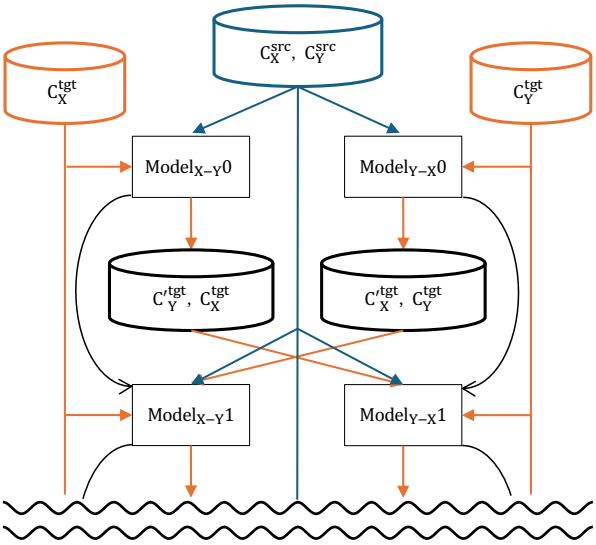

Figure 1: Iterative Back-translation process

predict words benefit most from Back-Translation. Since their analysis and method only consider words that appear in bilingual data, it is not clear whether words that appear only in monolingual resources can be translated. In addition, the method they investigated is Back-Translation, which is much less effective for acquisition of new words than IBT as we will show in a later section.

3. Iterative Back-Translation

Figure 1 illustrates Iterative Back-translation (IBT) process. Let X and Y denote two languages, then X-Y denote the translation from X to Y.

1. Train $Model_{X-Y}(0)$ and $Model_{Y-X}(0)$ using source domain bilingual data (C_X^{src},C_Y^{src}). Initialize i to 0.
2. Iterate the following steps:
   - 2.1 Translate target domain monolingual data C_Y^{tgt} by $Model_{Y-X}(i)$ and create pseudo bilingual data ($C_X'^{tgt}, C_Y^{tgt}$). Using ($C_X'^{tgt}, C_Y^{tgt}$) and (C_X^{src}, C_Y^{src}), fine-tune $Model_{X-Y}(i)$ into $Model_{X-Y}(i+1)$.
   - 2.2 Do the same on the opposite $Y-X$ direction. Set $i \leftarrow i+1$.

4. Research Method

In order to investigate how Acquirable Bilingual Vocabulary(ABV) is acquired through IBT, we first identify the target domain specific words (TDSWs) for both language X and Y by comparing source domain bilingual data (C_X^{src}, C_Y^{src}) and target domain monolingual data C_X^{tgt} and C_Y^{tgt}. Then, we evaluate translation models in terms of Acquisition Rate of Acquirable Bilingual Vocabulary (ARABV), which is defined on the test data by using the TDSWs. The process is described as follows.

1. The TDSWs D_X of the language X is identified from C_X^{tgt} and C_X^{src} as:

$$D_X = V(C_X^{tgt}) - V(C_X^{src})$$

   where $V(C)$ denotes the set of words in the data C. Likewise, D_Y = $V(C_Y^{tgt}) - V(C_Y^{src})$ is identified.

2. A word alignment tool is applied to the test data of the target domain $T = \{(s_X, s_Y)\}$ to obtain a set of aligned word pairs $A(s_X, s_Y) = \{(w_X, w_Y)\}$ for each sentence pair (s_X, s_Y). In this paper, we employed the word alignment tool provided in Moses (Koehn et al., 2007) [1].

3. For each word alignment $A(s_X, s_Y)$, we identify the ABV $B(s_X, s_Y)$, which is the word pairs of TDSWs:

$$B(s_X, s_Y) = \{(w_X, w_Y) | (w_X, w_Y) \in A(s_X, s_Y) \wedge w_X \in D_X \wedge w_Y \in D_Y\}$$

4. Our evaluation metric $ARABV(M_{X-Y})$ of a translation model M_{X-Y} is defined as follows.

$$ARABV(M_{X-Y}) = \frac{\sum_{(s_X,s_Y)\in T} \left| \begin{array}{r} \{w_X | (w_X, w_Y) \in B(s_X, s_Y) \\ \wedge w_Y \in M_{X-Y}(s_X)\} \end{array} \right|}{\sum_{(s_X,s_Y)\in T} |\{w_X | (w_X, -) \in B(s_X, s_Y)\}|}$$

   where $M_{X-Y}(s_X)$ is the translated sentence of language Y from s_X by the translation model M_{X-Y}.

We evaluate translation models in terms of ARABV and BLEU, a standard evaluation metric for MT.

5. Experiments

We investigate how ABV is acquired by domain adaptation using IBT by checking ARABV from a translation results of the translation model. We also investigate the impact of data preprocessing and differences in monolingual data on ARABV. To evaluate translation performance and acquisition of ABV, we use BLEU and ARABV respectively.

[1] In order to improve the accuracy of the word alignments, bilingual sentences of the target domain training data (C_X^{tgt}, C_Y^{tgt}) is also added to the training data of the word alignment.

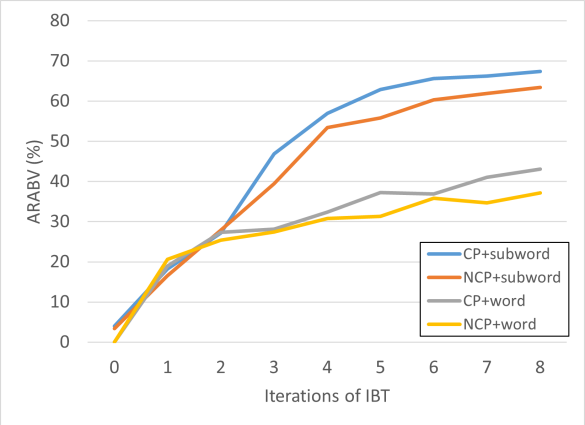

Figure 2: ARABV by IBT models of En-Ja translation direction

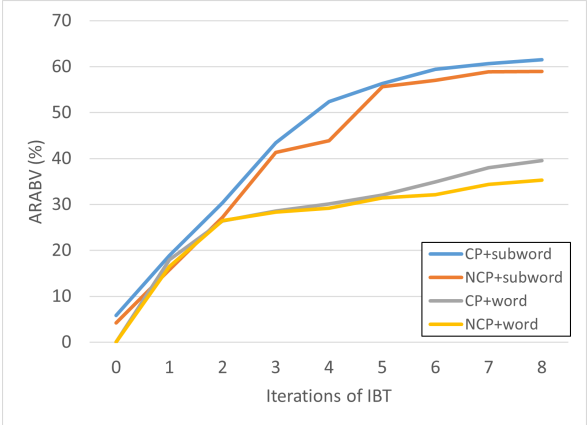

Figure 4: ARABV by IBT models of Ja-En translation direction

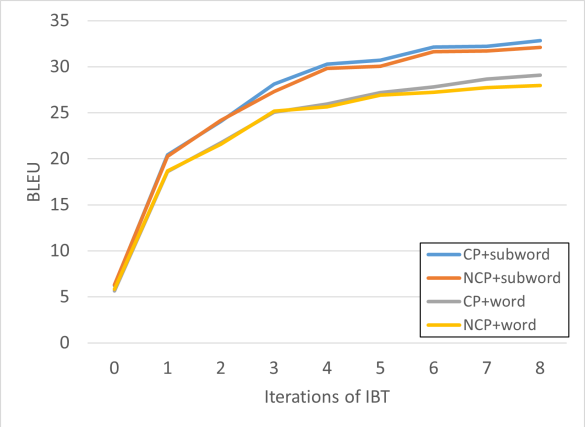

Figure 3: BLEU by IBT models of En-Ja translation direction

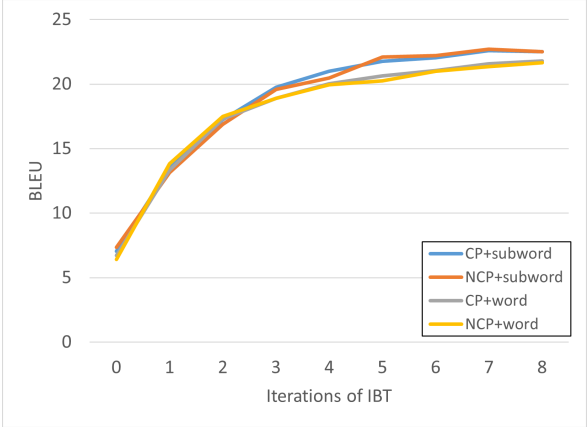

Figure 5: BLEU by IBT models of Ja-En translation direction

5.1. Dataset

We use the En-Jp dataset of The Kyoto Free Translation Task (KFTT) (Neubig, 2011) as source domain bilingual data and Asian Scientific Paper Excerpt Corpus (ASPEC) (Nakazawa et al., 2016) as target domain monolingual data. To use the ASPEC as monolingual data, 1M sentence pairs taken from the training data of the ASPEC are divided into two halves, each of which has 500k sentence pairs. From them, we created two pairs of monolingual data, CP and NCP. CP consists of the first halves of both languages, so that they compose a comparable data. NCP consists of the first half in English and the latter half in Japanese, so that they are non-comparable with each other. CP examples an ideal setting because ABV appears in the same context in the monolingual data of each language, while NCP examples a the realistic setting because ABV does not always appear in the same context. We found 806 and 796 pairs of ABV on the ASPEC test data with respect to CP and NCP, respectively.

In addition, we compared two different tokenization units, word and subword[2]. For the word tokenization, we employed the word-tokenizer provided in Moses toolkit (Koehn et al., 2007) in English and the morphological analyzer, Mecab (Kudo et al., 2004), in Japanese. For the subword tokenization, we employed SentencePiece (Kudo and Richardson, 2018).

5.2. Results

Figure 2, 3, 4 and 5 show experimental results of ARABV and BLEU in the En-Ja and Ja-En directions, respectively. The results show that repeating IBT process gradually increases ARABV and improves the translation performance (BLEU). Similar results are observed for Ja-En direction. These results suggest that IBT can acquire ABV,

[2]For the word setting, we put all words in ABV in the target vocabulary of NMT models so that they can output those in target sentences. On the other hand for the subword setting, we did not add any subwords for ABV.

a property that contributes to improved translation performance. When comparing CP and NCP setting, CP performed better but NCP still achieved good ARABV and BLEU. Several examples of correctly acquired non-trivial ABV are (dielectic, 誘電), (nonlinear, 非線形), (broadband, 広帯域), (superconducting, 超伝導), (diffraction, 回析), (perforation, 穿孔), (dialysis, 透析), (antibody, 抗体), (lesion, 病変), (ligament, 靭帯), etc.

Subword achieved a higher ARABV than word. This indicates the advantage of using subwords as a basic unit of MT. Basically, there's no way the initial model ($Model(0)$) can acquire any of ABV since it knows nothing about TDSWs. However, a small amount of ABV was correctly translated in the subword setting. This is because the transliteration type of ABV (e.g. albumin and アルブミン (*a-ru-bu-mi-n*, 'albumin'), UHV and UHV, etc.) can be easily translated by using subwords. On the other hand, the result also shows that IBT still improves ARABV even on the word setting. That indicates that the use of subwords is not a prerequisite for ABV acquisition.

6. Detailed Analysis

In this section, we conduct an analysis to explain why ABV is acquired through IBT. Hereinafter. we use experimental results in the subword setting. Firstly, we classify word pairs in ABV into four types by looking at the Japanese side as follows.

identical Japanese and English words are same.

kanji The Japanese word consists of only Kanji, Chinese characters. This type examples non-trivial bilingual words with each other.

katakana The Japanese word consists of only Katakana, Japanese phonograms. This type examples the Japanese and English words are transliteration with each other.

others Those other than any of above.

Table 1 shows examples of bilingual words for each type. We calculated ARABVs type by type. The following analysis is conducted for the first three types, excluding "others".

6.1. Word Frequency

To see how the number of occurrences of ABV in monolingual data affects ARABV, we grouped words in the ABV into bins according to their frequency in the monolingual data and examined ARABV for each group. Figure 6 shows the results of ARABV by type and number of occurrences in the monolingual data in the En-Ja direction.

This result shows that, in general, the more frequently the word appears, the higher its ARABV becomes. It also shows that the ABV of "identical" and "katakana" type is still acquired even if its frequency is low. That is because introduction of subword itself enables transliteration (Sennrich et al., 2016b).

Type	English	Japanese
identical	MIC	MIC
kanji	transfusion	輸血 (*yu-ke-tsu*, 'transfusion')
katakana	intranet	イントラネット (*i-n-to-ra-ne-Q-to*, 'intranet')
others	convulsion	けいれん (*ke-i-re-n*, 'convulsion')

Table 1: Examples of word pair type

ABV		Translation Result
Source	Target	
coaxial	同軸 (*dō-ji-ku*, 'coaxial')	光ファイバ (*hi-ka-ri-fa-i-ba*, 'optical fiber')
防臭 (*bō-shū*, 'deodorization')	deodorization	antibacterial

Table 2: examples of incorrect translation

6.2. ABV that cannot be Acquired

We examine what the ABV source word whose target word is not acquired through IBT is wrongly translated into. An example of the results is shown in Table 2. The word "coaxial" in the ABV in the table was often translated into "光ファイバ" (*hi-ka-ri-fa-i-ba*, 'optical fiber') instead of "同軸" (*dō-ji-ku*, 'coaxial'). This is because the contexts in which they appear are close, and "光ファイバ" appears more frequently in the training data. These results suggest that the context of the word plays a critical role in the success of the acquisition.

6.3. Experimet on different language setting

We also experiment on De-En datasets with CP+subword setting. As source and target domain data, we use News Commentary consisting of 201,288 sentences and Europarl consisting of 1,920,209 sentences of WMT14 Dataset, respectively. To increase the size of ABV, we combined the test data taken from WMT06, 07 and 08 to get that of 6K sentences in Europarl domain. We found 574 pairs of ABV on them.

Figure 7 and 8 show the results of ARABV and BLEU in each language direction. Even in this setting, ABV is acquired by IBT. Several examples of correctly acquired ABV are (Klaß, Klass),

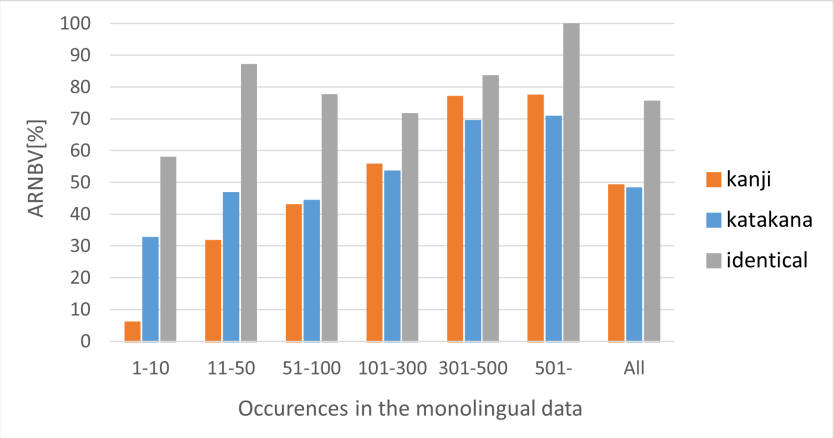

Figure 6: ARABV by number of occurrences in the monolingual data in the En-Ja direction

(audiovisuellen, audiovisual), (gütlich, civilised), (Mitentscheidung, codecision), (verehrte, honourable), (Bürgerbeauftragten, Ombudsman), (Supranationalität, supranationality), etc.

7. Discussion

The principle of the acquisition of a ABV by IBT can be explained as follows.

Suppose we have a word pair (x, y) in the ABV. The initial translation model $Model_{X-Y}(0)$ in IBT may be unable to translate x appeared in the monolingual corpus of language X correctly but still be able to translate its context, say $context(x)$, into language Y. Therefore, the created pseudo parallel corpus has pairs of sentences containing x (and $context(x)$) on X side and the translation of x's context, say $Trans_{X-Y}(context(x))$, on Y side. Then, it is to be used to train the opposite directional translation model $Model_{Y-X}(1)$, which may learn a translation rule from $Trans_{X-Y}(context(x))$ into x and $context(x)$. Since it is expected that $Trans_{X-Y}(context(x))$ is similar to $context(y)$, $Model_{X-Y}(1)$ has a chance to translate y and $context(y)$ appeared in the monolingual corpus of Y into x and $context(x)$. During IBT iterations, once $Model_{X-Y}(i)$ successfully translate y and $context(y)$ into x and $context(x)$, the next round of pseudo parallel corpus has that translation pair so that $Model_{Y-X}(i+1)$ may learn to translate y to x. Again, once $Model_{Y-X}$ successfully acquire that translation pair, it also contributes to train $Model_{X-Y}$ to translate x into y.

That hypothesis explains the gradual acquisition through IBT shown in Section 4.2. Section 5.1 demonstrated that words that are more frequent in monolingual data have a higher chance of being acquired. Additionally, Section 5.3 showed that ABV acquisition is more likely to occur for words with similar contextual usage. These findings support the our hypothesis on the acquisition process.

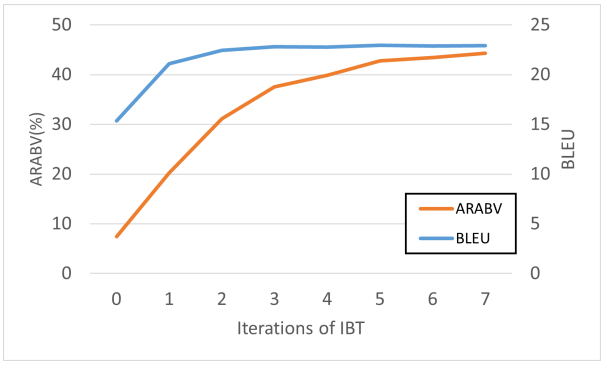

Figure 7: ARABV and BLEU by IBT models of En-De translation direction

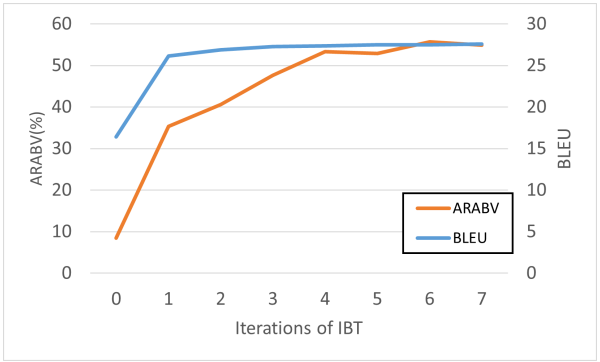

Figure 8: ARABV and BLEU by IBT models of De-En translation direction

8. Conclusion

In this paper, we revealed IBT can acquire ABV. In addition, the process and conditions of ABV acquisition were dusscussed. Our experimental results suggest that context of the word plays a critical role in the success of the acquisition.

9. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 23K11118.

10. References

Marzieh Fadaee and Christof Monz. 2018. Back-translation sampling by targeting difficult words in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 436–446, Brussels, Belgium. Association for Computational Linguistics.

Yinuo Guo, Hualei Zhu, Zeqi Lin, Bei Chen, Jian-Guang Lou, and Dongmei Zhang. 2021. Revisiting iterative back-translation from the perspective of compositional generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime Carbonell. 2019. Domain adaptation of neural machine translation by lexicon induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2989–3001, Florence, Italy. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Barcelona, Spain. Association for Computational Linguistics.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. ASPEC: Asian scientific paper excerpt corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2204–2208, Portorož, Slovenia. European Language Resources Association (ELRA).

Graham Neubig. 2011. The Kyoto free translation task. http://www.phontron.com/kftt.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. 2018. Joint training for neural machine translation models with monolingual data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).