Multi-Task Learning for Cross-Lingual Abstractive Summarization

Sho Takase, Naoaki Okazaki

Tokyo Institute of Technology
{sho.takase@nlp.c, okazaki@c}.titech.ac.jp

Abstract

We present a multi-task learning framework for cross-lingual abstractive summarization to augment training data. Recent studies constructed pseudo cross-lingual abstractive summarization data to train their neural encoder-decoders. Meanwhile, we introduce existing genuine data such as translation pairs and monolingual abstractive summarization data into training. Our proposed method, Transum, attaches a special token to the beginning of the input sentence to indicate the target task. The special token enables us to incorporate the genuine data into the training data easily. The experimental results show that Transum achieves better performance than the model trained with only pseudo cross-lingual summarization data. In addition, we achieve the top ROUGE score on Chinese-English and Arabic-English abstractive summarization. Moreover, Transum also has a positive effect on machine translation. Experimental results indicate that Transum improves the performance from the strong baseline, Transformer, in Chinese-English, Arabic-English, and English-Japanese translation datasets.

Keywords: summarization, machine translation, cross-lingual summarization

1. Introduction

Cross-lingual abstractive summarization is the task to generate a summary of a given document in a different target language (Leuski et al., 2003; Wan et al., 2010; Duan et al., 2019; Zhu et al., 2019). This task provides the overview of an article in a foreign language and thus helps readers understand a text written in an unfamiliar language quickly.

Early work on cross-lingual abstractive summarization adopted the pipeline approach: either translation of the given document into the target language followed by summarization of the translated document (Leuski et al., 2003) or summarization of the given document followed by translation of the summary into the target language (Orăsan and Chiorean, 2008; Wan et al., 2010). On the other hand, recent studies have applied a neural encoder-decoder model, which is widely used for natural language generation tasks including machine translation (Sutskever et al., 2014) and monolingual abstractive summarization (Rush et al., 2015), to generate a summary in the target language from the given document directly (Duan et al., 2019; Zhu et al., 2019). Such direct generation approaches prevent the error propagation in pipeline methods.

Training neural encoder-decoder models requires numerous sentence pairs. In fact, Rush et al. (2015) provided 3.8M sentence-summary pairs to train their neural encoder-decoder model for English abstractive summarization, and the following studies used the same training data (Zhou et al., 2017; Kiyono et al., 2017; Cao et al., 2018). However, constructing a large-scale cross-lingual abstractive summarization dataset is much more difficult than collecting monolingual summarization datasets because we require sentence-summary pairs in different languages. To address this issue, recent studies applied a machine translation model to monolingual sentence-summary pairs (Ayana et al., 2018; Duan et al., 2019; Zhu et al., 2019). They used the constructed pseudo dataset to train their neural encoder-decoder models.

Meanwhile, the possibility whether existing genuine parallel corpora such as translation pairs and monolingual abstractive summarization datasets can be utilized needs to be explored. In machine translation, Dong et al. (2015) indicated that using translation pairs in multiple languages improved the performance of a neural machine translation model. Similarly, we consider that such existing genuine parallel corpora have a positive influence on the cross-lingual abstractive summarization task since the task is a combination of machine translation and summarization.

In this study, we propose a multi-task learning framework, **Transum**, which includes machine translation, monolingual abstractive summarization, and cross-lingual abstractive summarization, for neural encoder-decoder models. The proposed method controls the target task with a special token which is inspired by Google's multilingual neural machine translation system (Johnson et al., 2017). For example, we attach the special token `<Trans>` to the beginning of the source-side input sentence in translation.

The proposed Transum is quite simple because it does not require any additional architecture in contrast to Zhu et al. (2019) but effective in cross-lingual abstractive summarization. Experimental results show that Transum improves the performance of cross-lingual abstractive summarization and outperforms previous methods in Chinese-English and Arabic-English summarization (Duan et al., 2019; Ouyang et al., 2019; Zhu et al., 2019; Zhu et al., 2020). In addition, Transum significantly improves machine translation performance compared to that obtained using only a genuine parallel corpus for machine translation.

Furthermore, we construct a new test set to simulate

more realistic situations: cross-lingual summarization with several length constraints. In a summarization process, it is important to generate a summary of a desired length (Takase and Okazaki, 2019). However, existing test sets for cross-lingual abstractive summarization cannot evaluate whether each model controls output lengths because the test sets do not contain summaries with multiple lengths. Thus, we translate an existing monolingual abstractive summarization that contains summaries with multiple lengths to construct the new test set.

The contributions of this study are as follows: 1. We propose a multi-task learning framework, **Transum**, that uses existing genuine parallel corpora in addition to pseudo cross-lingual abstractive summarization data to train a neural encoder-decoder model. 2. Transum achieves better scores than previous methods in Chinese-English and Arabic-English summarization. In addition, Transum also improves the performance of machine translation tasks. 3. We construct a new test set for cross-lingual summarization to simulate more realistic situations. The test set contains summaries with multiple lengths tied to one source document[1].

2. Task Definition

We begin by defining each task addressed in this study before describing our proposed framework.

2.1. Machine Translation

The purpose of machine translation is to generate a sentence in a target language from one in a source language while maintaining its content. Let V_s, V_t be the vocabulary of the source and target languages respectively, $\boldsymbol{x^s} = (\boldsymbol{x_1^s}, ..., \boldsymbol{x_M^s})$ be a given sentence in the source language ($\boldsymbol{x_i^s} \in \{0, 1\}^{|V_s|}$). We compute the following $\boldsymbol{y^t} = (\boldsymbol{y_1^t}, ..., \boldsymbol{y_N^t})$, which is the sentence in the target language ($\boldsymbol{y_j^t} \in \{0, 1\}^{|V_t|}$):

$$\arg \max_{\boldsymbol{y^t}} P(\boldsymbol{y^t}|\boldsymbol{x^s}). \quad (1)$$

2.2. Abstractive Summarization

The purpose of abstractive summarization is to generate a condensed summary of a given sentence. In addition, we have to preserve a length constraint for the generated summary in real applications (Takase and Okazaki, 2019). Thus, we generate a summary with the desired length L. Formally, we compute the following $\boldsymbol{y^s}$ for the given sentence $\boldsymbol{x^s}$ and desired length L:

$$\arg \max_{\boldsymbol{y^s}} P(\boldsymbol{y^s}|\boldsymbol{x^s}, L), \quad (2)$$

where the length of $\boldsymbol{y^s}$ is equal to L.

2.3. Cross-Lingual Abstractive Summarization

In cross-lingual abstractive summarization, we generate a summary in the target language with the desired length L for a given sentence. Thus, we compute the following $\boldsymbol{y^t}$:

$$\arg \max_{\boldsymbol{y^t}} P(\boldsymbol{y^t}|\boldsymbol{x^s}, L), \quad (3)$$

where the length of $\boldsymbol{y^t}$ is equal to L; this is the same as in (monolingual) abstractive summarization.

3. Proposed Method: Transum

Figure 1 shows an overview of the proposed multi-task learning framework: **Transum** which uses both of translation pairs and monolingual sentence-summary pairs for cross-lingual abstractive summarization. We attach the special token `<Trans>` or `<Summary>` to the beginning of the input sentence as shown in this figure. We expect to control the output of the decoder with the special token in a similar manner to that used in multilingual translation, where a special token indicates the target language (Johnson et al., 2017). Our used special token depends on only the target side. In other words, we use `<Summary>` for both of monolingual summarization and cross-lingual summarization. This configuration enables Transum to process cross-lingual abstractive summarization even if we train it on only translation pairs and monolingual summarization data (i.e., zero-shot generation). For example, if we replace `<Trans>` with `<Summary>` in the input Chinese sentence as shown in Figure 1, the decoder should generate a summary in English.

As described in Section 2, we have to preserve the length constraint in decoding abstractive summarization. Thus, we introduce a method to control output sequence length into our encoder-decoder. In this study, we adopt the length-ratio positional encoding (LRPE) (Takase and Okazaki, 2019), which is a variant of the sinusoidal positional encoding (Vaswani et al., 2017), to generate an output with the desired length L in Transformer. Let pos be the position of an input token and d be the embedding size. Then, the i-th dimension of LRPE, i.e., $LRPE_{(pos,L,i)}$ is as follows:

$$LRPE_{(pos,L,2i)} = \sin\left(\frac{pos}{L^{\frac{2i}{d}}}\right), \quad (4)$$

$$LRPE_{(pos,L,2i+1)} = \cos\left(\frac{pos}{L^{\frac{2i}{d}}}\right). \quad (5)$$

Following Takase and Okazaki (2019), we use LRPE for the decoder side during summarization. Thus, as shown in Figure 1, we adopt LRPE in the case that `<Summary>` is attached to the input sentence; otherwise, we use the conventional sinusoidal positional encoding.

---

[1]The constructed test set is publicly available at https://cl.asahi.com/api_data/jnc-jamul-en.html.
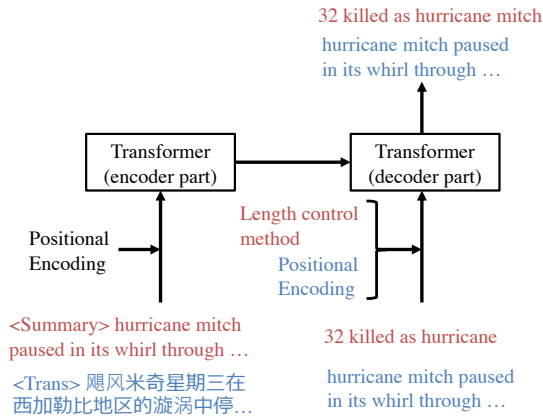
Figure 1: The overview of our proposed Transum. Transum requires attaching the special token `<Trans>` or `<Summary>` to the beginning of the input sentence and outputs the sequence based on the attached token. Thus, we can use both translation pairs and monolingual sentence-summary pairs in addition to the cross-lingual abstractive summarization data to train the neural encoder-decoder.

(a) Pseudo data construction from English sentence-summary pair

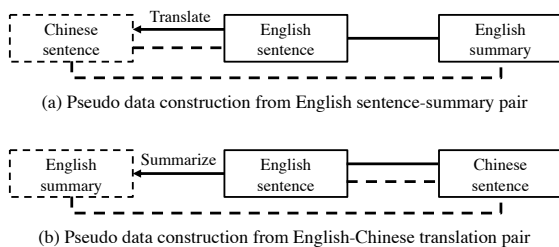(b) Pseudo data construction from English-Chinese translation pair

Figure 2: An example of pseudo data construction (Chinese-English abstractive summarization in this case) from parallel corpora. The solid line box denotes genuine data and the dashed line box denotes automatically generated pseudo data. The arrow indicates generation by the neural encoder-decoder model. The solid line denotes genuine parallel data and the dashed line denotes pseudo parallel data. In (a), we use the generated Chinese sentence for pseudo translation data in addition to pseudo Chinese-English summarization pairs. In (b), we also treat genuine translation pairs as cross-lingual summarization data without any compression.

4. Pseudo Data Construction

Duan et al. (2019) constructed pseudo cross-lingual abstractive summarization data from monolingual abstractive summarization data by back-translation (Sennrich et al., 2016). In this study, we also construct pseudo training data. In addition to monolingual summarization data, we use translation pairs as a source parallel corpus for pseudo data construction. Figure 2 shows an example of pseudo data construction for Chinese-English abstractive summarization.

4.1. From Monolingual Summarization

Figure 2 (a) illustrates the process of constructing pseudo Chinese-English summarization data from English sentence-summary pairs. Following Duan et al. (2019), we apply a neural machine translation model (English-to-Chinese in this figure) to the source sentences in the monolingual abstractive summarization dataset. Then, we use the pairs of translated sentences and genuine summaries as pseudo cross-lingual sentence-summary pairs.

In addition, we use the pairs of translated sentences and genuine source sentences as pseudo translation pairs. For these pairs, we attach a special token `<PseudoTrans>` to the source side sentence to indicate pseudo translation data such as Caswell et al. (2019).

4.2. From Translation Pairs

Figure 2 (b) illustrates the process of constructing pseudo Chinese-English summarization data from English-Chinese translation pairs. As shown in this figure, we apply a monolingual abstractive summarization model to sentences in the target language (English in this figure) to generate summaries. Then, we use the pairs of generated summaries and sentences in the source language as pseudo cross-lingual sentence-summary pairs. Moreover, we can consider the genuine translation pairs as the cross-lingual sentence-summary pairs without any compression. Thus, we also use the translation pairs as the cross-lingual sentence-summary pairs.

5. Experiments

We conduct experiments on three language pairs: Chinese-English, Arabic-English, and English-Japanese to evaluate the effectiveness of our proposed multi-task learning framework, Transum. As described in Section 3, we used Transformer (Vaswani et al., 2017) as the neural encoder-decoder model and LRPE (Takase and Okazaki, 2019) to control output sequence length[2]. We constructed vocabulary based on the unigram language model (Kudo, 2018)[3]. We set vocabulary size 32K for each language in each corpus.

5.1. Baselines

We compared the following methods with Trunsum to investigate the effect of our multi-task learning framework.

Translation We trained Transformer (Vaswani et al., 2017) on genuine translation pairs and then used it to translate the given sentence. In other words, this method outputs translated sentences without any compression.

Translation+Summarization We trained Transformer on the monolingual abstractive summarization dataset. Then we used it to summarize the sentence translated by the above Translation baseline. Thus, this

---

[2]https://github.com/takase/control-length
[3]https://github.com/google/sentencepiece

configuration represents the pipeline approach with the neural encoder-decoder.

Zero-shot   As described in Section 3, our introduced special tokens `<Trans>` and `<Summary>` enable the neural encoder-decoder to process cross-lingual abstractive summarization without the corresponding parallel corpus. Thus, we used only translation pairs and a monolingual abstractive summarization dataset for training.

Pseudo only   In contrast to the zero-shot configuration, we trained Transformer with the pseudo cross-lingual abstractive summarization data only. We constructed the pseudo data by applying the Translation and Summarization baselines to genuine parallel corpora.

Methods of recent studies   We trained the methods proposed in recent studies for cross-lingual abstractive summarization (Zhu et al., 2019; Zhu et al., 2020)[45] on our used training data. In particular, comparison with the method of Zhu et al. (2019) is important in evaluating our proposed method since their method also uses both of genuine and pseudo data for training. In addition, we describe the reported scores in previous studies for Chinese-English and Arabic-English summarization.

5.2. Dataset

We describe genuine parallel corpora which are used for training each method and pseudo data construction. Moreover, we describe test data for each language pair.

Chinese-English   We used 0.9M Chinese-English sentence pairs extracted from LDC corpora[6] as genuine translation pairs. For English abstractive summarization, we used 3.8M sentence-summary pairs extracted from Annotated English Gigaword (Napoles et al., 2012)[7] by the pre-processing script of Rush et al. (2015)[8]. The original script converts digits and infrequent words into the special token but we ignored this procedure in the same as the recent study (Takase and Okazaki, 2019).

For evaluation, we used the Chinese-English summarization dataset constructed by Duan et al. (2019). They manually translated the English source sentences in DUC 2004 task 1 data (Over et al., 2007) into Chinese. The evaluation set contains 500 Chinese source sentences and four kinds of English reference summaries for each Chinese sentence.

Arabic-English   We used 1.1M Arabic-English sentence pairs extracted from LDC corpora[9] as genuine

translation pairs. For English abstractive summarization, we used the same dataset as in the Chinese-English experiment.

To evaluate the performance, we used the DUC 2004 task 3 dataset (Over et al., 2007). The dataset contains 240 English source sentences translated from Arabic documents and four kinds of English reference summaries for each source sentence. We obtained the original Arabic sentences from the Agence France Presse portion of Arabic Gigaword[10], and then used the pairs of Arabic sentences and English summaries as the evaluation set.

English-Japanese   In the above two configurations, we used English as the target language whereas we are interested in the case where English is the source language. However, we could not find a cross-lingual abstractive summarization evaluation set containing English sentences as the source sentences with length constraints for each summary. To construct such an evaluation set, we manually translated the source sentences in a Japanese summarization dataset (Hitomi et al., 2019), which contains source sentences and three kinds of summaries depending on length constraints, into English. The constructed evaluation set contains 1,489 English sentences and three kinds of Japanese summaries for each English sentence.

For genuine English-Japanese translation pairs, we used the JIJI corpus[11] which contains 0.2M sentence pairs. Since this corpus is too small to train a neural encoder-decoder, we used JparaCrawl (Morishita et al., 2020)[12] to augment the training data for the Japanese-English translation model. For Japanese abstractive summarization, we used the Japanese News Corpus (Hitomi et al., 2019)[13], which contains 1.9M sentence-summary pairs.

5.3. Results of Cross-Lingual Abstractive Summarization

Tables 1, 2, and 3 show the recall-oriented ROUGE scores of each method in the Chinese-English, Arabic-English, and English-Japanese abstractive summarization test sets respectively. For Chinese-English and Arabic-English, we truncated characters over 75 bytes in generated summaries before computing ROUGE scores following the DUC 2004 evaluation protocol (Over et al., 2007). For English-Japanese, we truncated characters exceeding each desired length L before ROUGE computation. These tables contain ROUGE-1, 2, and L scores (R-1, R-2, and R-L respectively).

These tables indicate that the proposed Transum achieved the top score in each test set. These results show that our proposed multi-task learning framework

---

[4]https://github.com/ZNLP/NCLS-Corpora

[5]https://github.com/ZNLP/ATSum

[6]The corpora include LDC2004T07, Hansards portion of LDC2004T08, LDC2005T06, and LDC2015T06. We tried to obtain LDC2003E14 which is used in previous studies (Wang et al., 2017; Zhang et al., 2019) but it was unavailable because LDC has stopped its distribution. Thus, the number of translation pairs used in this study is less than the number used in previous studies.

[7]https://catalog.ldc.upenn.edu/LDC2012T21

[8]https://github.com/facebookarchive/NAMAS

[9]The corpora include LDC2004T17, LDC2004T18, LDC2005T05, and LDC2007T08.

[10]https://catalog.ldc.upenn.edu/LDC2011T11

[11]http://lotus.kuee.kyoto-u.ac.jp/WAT/jiji-corpus/

[12]http://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/

[13]https://cl.asahi.com/api_data/jnc-jamul-en.html

[15]We exclude the scores of (Cao et al., 2018) because they did not report recall-oriented ROUGE scores but F-1 ROUGE scores.

Method	R-1	R-2	R-L
Ayana et al. (2018)	19.3	4.3	17.0
Cheng et al. (2017)	24.0	7.0	21.3
Duan et al. (2019)	26.0	8.0	23.1
Translation	14.96	3.01	13.17
+ Summarization	18.82	5.08	17.00
Zero-shot	19.20	4.92	17.01
Pseudo only	24.72	7.37	21.29
Zhu et al. (2019)	19.92	6.08	17.91
Zhu et al. (2020)	18.74	5.46	17.03
Transum (proposed)	**26.69**	**8.49**	**23.33**

Table 1: Recall-oriented ROUGE scores of each method in Chinese-English abstractive summarization. The upper part of this table shows scores reported in previous studies[15].

Method	R-1	R-2	R-L
Doran et al. (2004)	25.87	4.73	21.98
Ouyang et al. (2019)	29.43	7.02	19.89
Translation	29.57	10.01	26.21
+ Summarization	31.85	10.98	28.48
Zero-shot	29.89	10.34	26.87
Pseudo only	35.78	11.98	31.30
Zhu et al. (2019)	31.38	11.29	28.35
Zhu et al. (2020)	29.43	10.33	26.75
Transum (proposed)	**36.04**	**12.12**	**31.49**

Table 2: Recall-oriented ROUGE scores of each method in Arabic-English abstractive summarization. The upper part of this table shows scores reported in previous studies.

had a positive influence on the cross-lingual abstractive summarization task. In particular, Transum outperformed Pseudo only approach. This result indicates that it is effective to use genuine translation pairs and monolingual summarization datasets.

The pseudo only approach achieved better performance than the zero-shot method in all language pairs. This result indicates that we should train the encoder-decoder with the dataset corresponding to the target task even if the training dataset is automatically constructed.

The zero-shot approach achieved comparable ROUGE scores to Translation except for Chinese-English and English-Japanese $L = 26$. This result implies that it is difficult to process the task in the zero-shot configuration. In contrast, the pipeline approach, which is a combination of machine translation and summarization, outperformed the zero-shot method in most test sets. Thus, it is better to combine the models trained for specific tasks than zero-shot in cross-lingual summarization.

Tables 1, 2, and 3 show that Transum outperformed the recent methods (Zhu et al., 2019; Zhu et al., 2020). However, these methods cannot control output sequence length and thus these comparisons might be unfavorable for them. Therefore, we compared these methods with
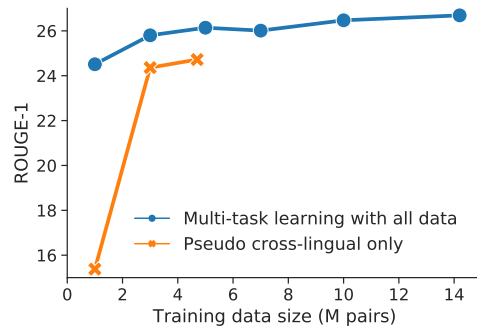

Figure 3: ROUGE-1 scores in the Chinese-English abstractive summarization test set. This figure reports the scores of models trained with varying the number of training pairs.

Transum without the length control method (LRPE) in an additional test set. We used the Chinese-English summarization test set constructed from Annotated English Gigaword (Napoles et al., 2012) by Duan et al. (2019). Table 4 shows F-1 based ROUGE scores of each method. This table indicates that Transum achieved better scores than those of Zhu et al. (2019) and Zhu et al. (2020). These results show that our multi-task learning framework offers a better approach than methods in these studies.

We investigate details of the contribution of each training data to the performance. Table 5 shows the performance in Chinese-English summarization in the case where we varied the kinds of datasets used to train the encoder-decoder. This table shows BLEU scores in the NIST 2002 Chinese-English translation test set and ROUGE-1 scores in the DUC 2004 task 1 (English abstractive summarizationz) in addition to ROUGE-1 scores in Chinese-English summarization. This table is composed of two parts: w/o and w/ constructed pseudo cross-lingual abstractive summarization data for training.

Table 5 indicates that the more training data we use, the better performance we achieve in Chinese-English translation and summarization. The upper part of this table shows that we improved ROUGE scores in Chinese-English abstractive summarization even though we did not use training data corresponding to this task. In particular, addition of pseudo translation pairs (4th row) raised more than four ROUGE-1 score compared to the case using only genuine translation pairs. These results imply that training with (pseudo) translation pairs has a positive effect on cross-lingual abstractive summarization. Moreover, we improved BLEU score more than eight points.

The lower part of Table 5 shows that the pseudo cross-lingual abstractive summarization dataset significantly improves ROUGE-1 scores in Chinese-English summarization. These results indicate that we need the dataset corresponding to the target task to achieve better scores. In contrast to other tasks, we cannot find significant

Method	$L = 10$			$L = 13$			$L = 26$		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Translation	9.31	1.41	8.89	13.16	1.91	11.87	18.88	2.81	14.51
+ Summarization	15.25	3.89	14.63	19.64	5.84	18.28	21.72	5.25	17.65
Zero-shot	11.40	2.82	11.14	14.76	3.61	13.77	24.61	5.84	19.17
Pseudo only	22.35	7.68	21.73	26.73	9.43	25.19	32.65	10.28	25.89
Zhu et al. (2019)	17.82	6.01	15.88	22.38	7.66	19.41	27.71	9.85	23.23
Zhu et al. (2020)	16.71	5.22	14.93	20.41	6.72	17.89	23.85	8.13	20.41
Transum (proposed)	**27.17**	**10.92**	**26.44**	**32.87**	**13.75**	**30.87**	**38.83**	**14.45**	**30.99**

Table 3: Recall-oriented ROUGE scores of each method in English-Japanese abstractive summarization. The English-Japanese test set contains three kinds of Japanese summaries depending on the desired length $L = 10, 13$, and 26. Thus, we report the scores for each desired length.

Method	R-1	R-2	R-L
Ayana et al. (2018)	21.5	6.6	19.6
Cheng et al. (2017)	26.7	10.2	24.3
Duan et al. (2019)	30.1	12.2	27.7
Cao et al. (2018)	**32.04**	13.60	27.91
Zhu et al. (2019)	29.61	12.55	27.41
Zhu et al. (2020)	28.19	11.89	26.28
Transum w/o LRPE	31.43	**13.99**	**29.04**

Table 4: The full-length F-1 based ROUGE scores of each method in the Chinese-English abstractive summarization dataset constructed from English Gigaword. The upper part of this table shows scores reported in previous studies.

improvement for ROUGE-1 scores in the monolingual abstractive summarization task. In other words, our multi-task learning framework had little effect on monolingual abstractive summarization. We consider the reason is that we did not prepare an additional monolingual dataset for training. Thus, we might improve its performance by adding any monolingual task such as paraphrasing.

Figure 3 illustrates ROUGE-1 scores in the Chinese-English summarization dataset when we varied the training data size. This figure contains two configurations: using the pseudo cross-lingual abstractive summarization data only and using both genuine and pseudo data. This figure shows that we can achieve better performance in using both genuine and pseudo data to train the model. This result indicates that multi-task learning is a better approach than using the training data corresponding to the target task only.

The ROUGE-1 score of the model trained with pseudo cross-lingual only is low when the training data size is small. Since this dataset contains only summarization, the number of tokens for generation is smaller than that in translation pairs. Thus, it might be difficult to learn the alignment of tokens in a source sentence with output tokens.

5.4. Results of Machine Translation

Since Table 5 shows that our Transum also improved the BLEU score in Chinese-English translation, we fur-

ther investigate the performance of Transum in machine translation. For Chinese-English and Arabic-English, we used NIST 2002-2005 test sets. We computed BLEU scores with the official tool. Moreover, we used the test set of the JIJI corpus for English-Japanese.

Tables 6, 7, and 8 show BLEU scores of Transformer trained with only genuine translation pairs (i.e., the Translation baseline) and our Transum in Chinese-English, Arabic-English and English-Japanese respectively. Each table also indicates BLEU scores reported in previous studies[16].

These tables show that Transum outperformed Transformer trained with only translation pairs except for MT03 in Arabic-English. Thus, our multi-task learning approach also has a positive effect on machine translation. In particular, Transum raised the BLEU score approximately eight or more points compared to the Transformer trained with only translation pairs in all test sets of Chinese-English translation. Moreover, Transum outperformed the previous top score on MT02 and MT03 in Chinese-English and JIJI English-Japanese translation even though Transum has no specific approach to machine translation. Furthermore, since the methods proposed in Zhang et al. (2019) and Cheng et al. (2019) are orthogonal to Transum, we expect further improvement in machine translation by introducing their methods into Transum.

6. Related Work

Early explorations on cross-lingual summarization adopted the pipeline approach, which combines the machine translation with summarization methods (Leuski et al., 2003; Orăsan and Chiorean, 2008; Wan et al., 2010; Yao et al., 2015). Orăsan and Chiorean (2008) applied the Maximal Marginal Relevance to summarize given documents and then automatically translated the summary. To prevent unreadable outputs, Wan et al. (2010) proposed the method to predict the machine translation quality for sentences in a source document, and then generate a summary based on the predicted

---

[16] Each table includes scores of single models. For Alrajeh (2018), we focused on the setting which uses almost the same translation pairs as ours.

Training data						MT02	DUC	
Genuine		Pseudo (cross-lingual sum)			Pseudo	Zh-En	En-En	Zh-En
Trans	Sum	From trans	From sum	Trans as sum	(trans)	BLEU	R-1	R-1
Cross-lingual abstractive summarization with zero-shot								
✓						40.86	-	14.96
✓	✓					41.87	30.36	17.77
✓	✓			✓		43.80	30.16	17.40
✓	✓				✓	47.87	29.97	19.20
✓	✓			✓	✓	**49.31**	30.11	18.06
Trained with pseudo cross-lingual abstractive summarization dataset								
✓	✓	✓				43.99	29.16	24.68
✓	✓	✓	✓			44.48	29.82	25.42
✓	✓	✓	✓	✓		45.26	29.68	26.05
✓	✓	✓	✓	✓	✓	48.64	**30.37**	**26.69**

Table 5: Results of the model trained with each training dataset. This table shows BLEU scores in Chinese-English (Zh-En) MT02 and ROUGE-1 scores in the original DUC 2004 task 1 and Chinese-English DUC 2004.

Method	MT02	MT03	MT04	MT05
Wang et al. (2017)	-	39.35	41.15	38.07
Cheng et al. (2018)	46.10	44.07	45.61	44.06
Cheng et al. (2019)	48.13	47.83	49.13	**49.04**
Zhang et al. (2019)	-	48.31	**49.40**	48.72
Transformer	40.86	40.71	40.51	37.22
Transum (proposed)	**48.64**	**50.23**	48.45	47.84

Table 6: BLEU scores of each method in the NIST Chinese-English test set.

Method	MT02	MT03	MT04	MT05
Almahairi et al. (2016)	-	-	-	51.19
Alrajeh (2018)	-	-	-	58.35
Transformer	50.54	**62.37**	50.89	52.96
Transum (proposed)	**53.34**	60.94	**55.53**	**58.57**

Table 7: BLEU scores of each method in the NIST Arabic-English test set.

Method	BLEU
Morishita et al. (2017)	19.41
Susanto et al. (2019)	21.91
Transformer	17.62
Transum (proposed)	**23.10**

Table 8: BLEU scores of each method in the JIJI English-Japanese test set.

quality score before translation. Yao et al. (2015) extended phrase-based machine translation models to select important phrases. However, Wan (2011) indicated that we should use information from both sides rather than such pipeline approaches.

Recent studies applied a neural encoder-decoder model to generate cross-lingual abstractive summaries from the given document directly (Ayana et al., 2018; Duan et al., 2019; Zhu et al., 2019). To construct pseudo cross-lingual abstractive summarization data for training, Ayana et al. (2018) adopted the approach consisting of two steps: machine translation of a document in the source language and then monolingual abstractive summarization of the translated document. Then, they used the pairs of summarized translations and the original documents as pseudo training data. Duan et al. (2019) used genuine summaries to improve the quality of constructed training data. They applied a machine translation model to source sentences in monolingual sentence-summary pairs and used the pairs of translated sentences and genuine summaries as pseudo training data. Zhu et al. (2019) proposed a round-trip translation strategy to obtain high quality pseudo training data from existing monolingual summarization datasets. Their round-trip strategy translates a source sentence in monolingual sentence-summary pairs in the same manner as Duan et al. (2019), and then re-translates the translated sentence into the source language. Their approach filters out based on the similarity between the source sentence and round-trip translation. These studies explored the sophisticated way to construct pseudo training data but pay little attention to the existing genuine parallel corpora. In contrast, this study utilizes such genuine parallel corpora in addition to constructed pseudo data with the multi-task learning framework.

In addition to the round-trip translation strategy, Zhu et al. (2019) introduced a multi-task learning approach for cross-lingual abstractive summarization. Their method prepares two decoders: one for cross-lingual abstractive summarization and the other for machine translation or monolingual summarization. The method trains the decoders to generate corresponding output for a given sentence. They indicated that their multi-task learning approach improved the performance of cross-lingual abstractive summarization but it requires additional parameters for a decoder. In contrast, Transum is more simple because it only needs to attach the special token to the source sentence. Moreover, experimental results show that Transum outperformed the multi-task approach of Zhu et al. (2019).

7. Conclusion

This paper presents a multi-task learning framework for cross-lingual abstractive summarization to augment training data. The proposed method, Transum, attaches the special token to the beginning of the input sentence to indicate the target task. The special token enables us to use genuine translation pairs and the monolingual abstractive summarization dataset in addition to the pseudo cross-lingual abstractive summarization data for training. The experimental results show that Transum achieved better performance than the pipeline approach and model trained with pseudo data only. We achieved the top ROUGE scores in Chinese-English and Arabic-English summarization. Moreover, Transum also improved the performance of machine translation and outperformed the previous top score in the JIJI English-Japanese translation.

Acknowledgements

References

Almahairi, A., Cho, K., Habash, N., and Courville, A. C. (2016). First result on arabic neural machine translation. *CoRR*, abs/1606.02680.

Alrajeh, A. (2018). A recipe for arabic-english neural machine translation. *CoRR*, abs/1808.06116.

Ayana, qi Shen, S., Chen, Y., Yang, C., yuan Liu, Z., and song Sun, M. (2018). Zero-shot cross-lingual neural headline generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(12):2319–2327.

Cao, Z., Wei, F., Li, W., and Li, S. (2018). Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, pages 4784–4791.

Caswell, I., Chelba, C., and Grangier, D. (2019). Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (WMT)*, pages 53–63.

Cheng, Y., Yang, Q., Liu, Y., Sun, M., and Xu, W. (2017). Joint training for pivot-based neural machine translation. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3974–3980.

Cheng, Y., Tu, Z., Meng, F., Zhai, J., and Liu, Y. (2018). Towards robust neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1756–1766.

Cheng, Y., Jiang, L., and Macherey, W. (2019). Robust neural machine translation with doubly adversarial inputs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4324–4333.

Dong, D., Wu, H., He, W., Yu, D., and Wang, H. (2015). Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (EMNLP)*, pages 1723–1732.

Doran, W., Stokes, N., Newman, E., Dunnion, J., Carthy, J., and Toolan, F. (2004). News story gisting at university college dublin. In *Proceedings of the HLT-NAACL 2004 Document Understanding Workshop*.

Duan, X., Yin, M., Zhang, M., Chen, B., and Luo, W. (2019). Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3162–3172.

Hitomi, Y., Taguchi, Y., Tamori, H., Kikuta, K., Nishitoba, J., Okazaki, N., Inui, K., and Okumura, M. (2019). A large-scale multi-length headline corpus for analyzing length-constrained headline generation model evaluation. In *Proceedings of the 12th International Conference on Natural Language Generation (INLG)*, pages 333–343.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics (TACL)*, 5:339–351.

Kiyono, S., Takase, S., Suzuki, J., Okazaki, N., Inui, K., and Nagata, M. (2017). Source-side prediction for neural headline generation. *CoRR*.

Kudo, T. (2018). Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 66–75.

Leuski, A., Lin, C.-Y., Zhou, L., Germann, U., Och, F. J., and Hovy, E. (2003). Cross-lingual c*st*rd: English access to hindi information.

Morishita, M., Suzuki, J., and Nagata, M. (2017). NTT neural machine translation systems at WAT 2017. In *Proceedings of the 4th Workshop on Asian Translation (WAT)*, pages 89–94.

Morishita, M., Suzuki, J., and Nagata, M. (2020). JParaCrawl: A large scale web-based English-Japanese parallel corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 3603–3609.

Napoles, C., Gormley, M., and Van Durme, B. (2012). Annotated Gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, AKBC-WEKEX, pages 95–100.

Orăsan, C. and Chiorean, O. A. (2008). Evaluation of a cross-lingual Romanian-English multi-document summariser. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*.

Ouyang, J., Song, B., and McKeown, K. (2019). A robust abstractive system for cross-lingual summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2025–2031.

Over, P., Dang, H., and Harman, D. (2007). Duc in context. *Information Processing & Management*, 43(6):1506–1520.

Rush, A. M., Chopra, S., and Weston, J. (2015). A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 379–389.

Sennrich, R., Haddow, B., and Birch, A. (2016). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 86–96.

Susanto, R. H., Htun, O., and Tan, L. (2019). Sarah's participation in WAT 2019. In *Proceedings of the 6th Workshop on Asian Translation (WAT)*, pages 152–158.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3104–3112.

Takase, S. and Okazaki, N. (2019). Positional encoding to control output sequence length. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 3999–4004.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5998–6008.

Wan, X., Li, H., and Xiao, J. (2010). Cross-language document summarization based on machine translation quality prediction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 917–926.

Wan, X. (2011). Using bilingual information for cross-language document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 1546–1555.

Wang, M., Lu, Z., Zhou, J., and Liu, Q. (2017). Deep neural machine translation with linear associative unit. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 136–145.

Yao, J.-g., Wan, X., and Xiao, J. (2015). Phrase-based compressive cross-language summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 118–127.

Zhang, W., Feng, Y., Meng, F., You, D., and Liu, Q. (2019). Bridging the gap between training and inference for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4334–4343.

Zhou, Q., Yang, N., Wei, F., and Zhou, M. (2017). Selective encoding for abstractive sentence summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1095–1104.

Zhu, J., Wang, Q., Wang, Y., Zhou, Y., Zhang, J., Wang, S., and Zong, C. (2019). NCLS: Neural cross-lingual summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3052–3062.

Zhu, J., Zhou, Y., Zhang, J., and Zong, C. (2020). Attend, translate and summarize: An efficient method for neural cross-lingual summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1309–1321.