Multilingual Parallel Corpus for Global Communication Plan

Kenji Imamura and Eiichiro Sumita
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
{kenji.imamura,eiichiro.sumita}@nict.go.jp

Abstract

In this paper, we introduce the Global Communication Plan (GCP) Corpus, a multilingual parallel corpus being developed as part of the GCP. The GCP Corpus is intended to be develop speech translation systems; thus, it primarily consists of pseudo-dialogues between foreign visitors and local Japanese people. The GCP Corpus is sentence-aligned and covers four domains and ten languages, including many Asian languages. In this paper, we summarize the GCP and the current status of the GCP Corpus. Then, we describe some of the corpus' basic characteristics from the perspective of multilingual machine translation and compare direct, pivot, and zero-shot translation techniques.

Keywords: Multilingual Parallel Corpus, Global Communication Plan, Asian Languages, Pivot/Zero-shot Translation

1. Introduction

In 2014, the Ministry of Internal Affairs and Communications in Japan implemented the Global Communication Plan (GCP; c.f., Sec. 2.), whose mission is to eliminate global "language barriers." An important aspect of this plan is the public demonstration of multilingual speech translation technologies. With this in mind, we are improving the usability of the multilingual speech translation system by improving translation quality and developing user interfaces that enable it to be used in various places, such as hospitals, malls/stores, and tourist spots.

Multilingual translation corpora are required to develop multilingual speech translation systems, and the GCP includes a corpus development program for machine translation (MT).

In this paper, we summarize the GCP and the current status of the GCP Corpus, which is a multilingual parallel corpus used for public demonstration [1]. This corpus primarily consists of pseudo-dialogues between foreign visitors and local Japanese people. (It also includes some isolated utterances, such as phrases frequently associated with travelling.) The utterances in the dialogues are translated into ten languages (including Japanese) targeted by the GCP. Therefore, the corpus is sentence-aligned. The target domains are medical care, disaster prevention, shopping, and tourism.

Although the GCP Corpus is being developed for use in speech translation systems, it could also be used in various other research fields because it has the following characteristics.

1. It is a multilingual sentence-aligned corpus that covers ten languages, including Asian languages. Therefore, this allows 90 different MT systems to be constructed. Furthermore, it can also be applied in comparative studies of pivot translation (Utiyama and Isahara, 2007; Cohn and Lapata, 2007) and zero-shot translation (Johnson et al., 2016).
2. It covers four domains, namely, medical care, disaster prevention, shopping, and tourism. It can be applied to domain adaptation studies (e.g., (Imamura and Sumita, 2016)).
3. It consists of pseudo-dialogues. Therefore, it can also be applied to discourse studies that consider long-distance contexts. Note that such contexts are simpler than those of real dialogues because the dialogue never breaks down (Higashinaka et al., 2016).

In this paper, we focus on the first characteristic. We use the GCP Corpus to confirm MT qualities between Japanese and other languages. Furthermore, we compare the qualities of direct, pivot, and zero-shot translations.

Europarl (Koehn, 2005), a collection of European Parliament proceedings, is a well-known multilingual parallel corpus. The characteristics of the GCP Corpus are similar to those of the Europarl. However, the GCP Corpus has different applications because it includes Asian languages and pseudo-dialogues that are being developed for use in speech translation systems.

The reminder of this paper is organized as follows. Sections 2. and 3. summarize the GCP and discuss the current status of the GCP Corpus, respectively. In Section 4., we construct a neural machine translation system using the GCP Corpus and evaluate the quality of its translations. In Section 5., we compare direct, pivot, and zero-shot translations. We discuss current problems and future directions in Section 6. and present our conclusions in Section 7.

2. Global Communication Plan

Global Communication Plan was proposed in 2014 by Yoshitaka Shindo, who was Minister of Internal Affairs and Communications of Japan. The mission of the GCP is to eliminate global "language barriers" by targeting the following goals.

1. Realizing global and open communications
2. Enhancing Japanese presence in the world
3. Promoting "O-mo-te-na-shi" (hospitality) at the Tokyo 2020 Olympic and Paralympic Games

[1] We are also developing speech corpora for speech recognition and synthesis tasks; however, we only describe the parallel corpus in this paper.

Category				
Lv. 1	Lv. 2	Lv. 3	Speaker Type	Utterance
Medical Care	Illness or Injury	Response to Urgency	Foreigner	Is there a hospital nearby?
			Japanese	There is an internal medicine hospital after you turn right at that convenience store.
			Japanese	Aren't you feeling well?
			Foreigner	I feel dizzy.
			Japanese	Aren't you good at Japanese?
			Foreigner	I am not good at it.
			Japanese	It may be impossible to communicate in English in the hospital there.
			Japanese	There are staffs who can speak English in the comprehensive hospital in front of the station nearby.

Table 1: Example Pseudo-dialogue for Medical Care Domain

Based on this plan, the Global Communication Developers' Group [2] was formed in collaboration with industry, academia, and the government. This group performs the following tasks: 1) research and development of multilingual speech translation to extend target languages and domains, and 2) public demonstrations of speech translation at hospitals, malls/stores, and tourist spots.
The GCP focuses on the following four domains.

- **Medical care**
  This domain includes dialogues between patients and medical staff (doctors, nurses, etc.) at hospitals.
- **Disaster prevention**
  This domain involves local governments dealing with disasters and providing information to foreigners.
- **Shopping**
  This domain includes dialogues between store clerks and foreign visitors who are shopping.
- **Tourism**
  This domain includes dialogues that provide travel information to visitors, e.g., information about accommodations, transportation, and tourist spots.

The ten target languages are as follows: Japanese (Ja), English (En), Simplified Chinese (Zh), Korean (Ko), Thai (Th), Vietnamese (Vi), Indonesian (Id), Myanmar (My), Spanish (Es), and French (Fr). These languages were selected taking into account the number of visitors to Japan who spoke these languages.

3. GCP Corpus

The GCP Corpus is being developed to help create speech translation systems for the GCP. In this paper, we focus on the translation component rather than the speech component.
The corpus targets four domains defined by the GCP. We assume situations where foreign visitors are speaking with local Japanese people because the goal of the corpus is to realize speech translation in such situations. Therefore, the corpus primarily consists of dialogues. However, these are not real dialogues that have been recorded and transcribed but pseudo-dialogues written by scenario writers imagining possible situations. Pseudo-dialogues are more suitable from the perspective of early development of speech translation systems because actual dialogues contain many ungrammatical utterances that require cleaning during transcription and are difficult for human translators to understand.
Table 1 shows an example pseudo-dialogue in the medical care domain. Here, each dialogue is categorized into three levels, and each utterance contains speaker type information (Japanese or foreigner). We also include isolated utterances, such as greetings and common expressions.
Since the GCP Corpus involves conversations between Japanese people and foreign visitors, we first created pseudo-dialogues in Japanese and then translated them into the other nine target languages. Table 2 shows the size of the corpora for each language at the end of 2017. The Japanese, English, Chinese, and Korean corpora are larger than the other languages because these languages are given first preference.

4. Quality of MTs Based on GCP Corpus

The GCP Corpus is being developed to help realize speech translation systems. In this section, we evaluate MT quality by training a neural MT (NMT) system using the GCP Corpus.

4.1. Experimental Settings

Language Pairs The GCP Corpus is a parallel corpus and includes ten languages; therefore, we can construct up to $10 \times 9 = 90$ different MT systems. Here, we only evaluate the quality of MTs between Japanese and the other languages (Ja $\leftrightarrow$ X; a total of 18 systems) due to resource limitations.

Datasets The corpora (Table 2) were divided into training, development, and test sets. Initially, we set aside some sentences from each corpus (held-out data) and used the remaining sentences as the training set. From the held-out data, we uniformly selected two 2,000 sentence sets as development and test sets.

MT System The training, development, and test sets were segmented into words using in-house word segmenters, and the words were further segmented into subwords using a byte-pair encoder (Sennrich et al., 2016).

[2] http://gcp.nict.go.jp/about/index.html

Language	Abbr.	No. of Sentences (Utterances) Total	Medical Care	Disaster Prevention	Shopping	Tourism	Other
Japanese	Ja	2,029,111 (25.2 chars. / sent.)	420,270	249,495	355,429	527,056	476,861
English	En	2,029,111 (11.2 words / sent.)	420,270	249,495	355,429	527,056	476,861
Chinese	Zh	2,026,608	420,270	249,495	355,429	527,056	474,358
Korean	Ko	2,026,608	420,270	249,495	355,429	527,056	474,358
Thai	Th	1,150,070	145,054	117,636	180,843	232,179	474,358
Vietnamese	Vi	1,150,070	145,054	117,636	180,843	232,179	474,358
Indonesian	Id	1,150,070	145,054	117,636	180,843	232,179	474,358
Myanmar	My	1,150,070	145,054	117,636	180,843	232,179	474,358
Spanish	Es	337,654	145,054	117,636	9,512	18,944	46,508
French	Fr	340,499	145,054	117,636	9,867	19,593	48,349

Table 2: GCP Corpora Sizes as of Summer 2017

The number of sub-word types (corresponding to the vocabulary size) was approximately 16 thousand per language.

We used OpenNMT (Klein et al., 2017) [3] as the neural translation system with the following settings.

- We used a two-layer bi-directional LSTM (long short-term memory) encoder with $500+500$ units. The word embedding was 500 units.
- We used a two-layer LSTM decoder with 1,000 units. The word embedding was 500 units.
- The stochastic gradient descent (SGD) was used for optimization with a learning rate of 1.0 for the first fourteen epochs, followed by annealing of six epochs while decreasing the learning rate by half. The mini-batch size was 64.
- During translation, 10-best translation was performed with a beam width of 10. Furthermore, we applied reranking using the following formula and selected the best translation (Morishita et al., 2017; Oda et al., 2017).

$$ll_{\mathrm{len}}(\boldsymbol{y}|\boldsymbol{x}) = \sum_t \log Pr(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t}) + WP \cdot T, \quad (1)$$

where ll_{len} denotes the log-likelihood that considers translation length, the first term of the right side denotes the log-likelihood, WP denotes a word penalty ($WP \geq 0$), and T denotes the word number of the translation.

Equation 1 corrects a translation length using the word penalty because NMTs typically generate short translations. The word penalty is optimized using a development set to make the translation length and reference length nearly equal. By correcting the translation length, we can compute the BLEU scores regardless of the brevity penalty.

4.2. Translation Quality

Table 3 shows the quality of the MTs as measured by BLEU (Papineni et al., 2002).

[3] http://opennmt.net/

Language	No. of Training Sentences	BLEU Score from Japanese	BLEU Score to Japanese
English	1,954,477	27.20	31.01
Chinese	1,952,475	35.82	42.34
Korean	1,952,475	52.87	58.13
Thai	1,110,232	25.64	27.64
Vietnamese	1,110,232	30.32	30.64
Indonesian	1,110,232	22.16	25.94
Myanmar	1,110,232	23.90	30.82
Spanish	326,433	22.28	24.82
French	329,160	22.05	23.39

Table 3: Quality of MTs between Japanese and other Languages

Languages	Ja → En	En → Ja
General	27.20	31.01
Medical Care	26.53	30.69
Disaster Prevention	27.23	33.01
Shopping	26.09	29.96
Tourism	32.47	33.22
Other	23.89	27.05

Table 4: Translation Quality of Each Domain

First, the BLEU scores are significantly different for each language, ranging from 22.05 to 52.87 for translation from Japanese and from 23.39 to 58.13 for translation to Japanese. However, the score tended to increase with an increasing number of training sentences.

Next, by comparing translations from Japanese with translations to Japanese, it was found that the scores when translating to Japanese were higher than those when translating from Japanese for all language pairs. This phenomenon shows that translating from Japanese was more difficult than translating to Japanese. For example, the subjects of Japanese sentences are sometimes not present, and MT systems translating to other languages must generate such subjects.

In addition, only the BLEU scores from/to Korean were greater than 50. It is known that conventional statistical MTs between Japanese and Korean tend to be high quality because Korean grammar is similar to Japanese grammar (e.g., SVO order). We observed similar results using NMT.

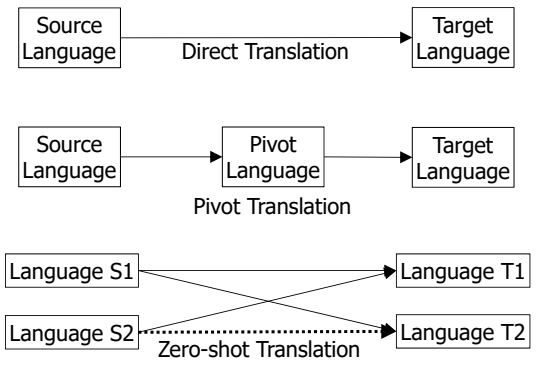

Figure 1: Direct, Pivot, and Zero-shot Translations

Table 4 shows the translation quality of each domain between Japanese and English. 'General' represents the scores of the general test sets described in Section 4.1. The test sets for the domains were extracted by selecting 1,000 sentences from the held-out data described in Section 4.1. The same models used in Table 3 were used for all domains (i.e., no domain adaptation was employed). This table shows that BLEU scores that are comparable to the general test sets were obtained in the target domains of GCP, although there are some variations.

5. Pivot and Zero-shot Translation

We can directly construct translators between all possible language pairs from multilingual parallel corpora (i.e., direct translation). However, if we do not have such parallel corpora, we use pivot translation, which involves translating source sentences into target sentences via a resource-rich language known as a pivot (Utiyama and Isahara, 2007; Cohn and Lapata, 2007).

In NMT, another approach known as the zero-shot translation can be used to construct MT systems without using directly-translated corpora (Johnson et al., 2016). Here, an encoder that recognizes the source language and a decoder that generates the target language are trained using the corpora of indirect language pairs. Figure 1 illustrates the relationship among these three approaches.

Both the pivot and zero-shot translation generally assume that bilingual corpora covering the source language and those that cover the target language are obtained from the different texts. However, comparative analysis is difficult to perform under this setting because the vocabulary differs and we cannot construct a direct translator.

Using multilingual parallel corpora, we can compare these three translation methods. By regarding the quality of the direct translation as the upper bound, we can evaluate improvements in pivot and zero-shot translations by comparing them to the direct translation. Note that the experimental settings can be shared for all three methods; for example, identical vocabularies can be used.

In this section, we compare direct, pivot, and zero-shot translation using four languages, i.e., Japanese, English, Chinese, and Korean.

Figure 2: Training Data for Zero-shot Translation (Ja-En)

5.1. Experimental Settings

Data From approximately two million sentences that were common among the four languages, we created training, development, and test sets using the same method in Section 4.1. The sentences were segmented into fifty thousand sub-word types using byte-pair encoding trained from all training sentences (i.e., joint encoding). Note that the same vocabulary set was used for all experiments discussed in this section.

Direct Translation Models for the 12 language pairs were trained using the system described in Section 4.

Pivot Translation The qualities of six translations among English, Chinese, and Korean were measured using Japanese as the pivot language. In other words, we measured the BLEU scores for (En | Zh | Ko) $\rightarrow$ Ja $\rightarrow$ (En | Zh | Ko) using the Ja $\leftrightarrow$ (En | Zh | Ko) models trained for direct translation tasks. In each case we used the 1-best translations from the source language to the pivot languages.

Zero-shot Translation We constructed (En | Zh | Ko) $\rightarrow$ (En | Zh | Ko) translators using the Ja $\leftrightarrow$ (En | Zh | Ko) corpora similar to the pivot translation.

First, we added a target language tag at the beginning of each source sentence in the Ja $\leftrightarrow$ (En | Zh | Ko) corpora (Figure 2). Then, a unified model was trained using a combined corpus containing all these language pairs. In this experiment, 12 million sentences were used for training.

During testing, the target language tags were added to the source sides of the test sentences, and translation was performed using the unified model. Then, appropriate target sentences were generated based on the tags even though that particular language pair had not been learned.

5.2. Results

Table 5 shows the results of the (a) direct, (b) pivot, and (c) zero-shot translations. With the pivot translations, the BLEU scores for most language pairs were worse than those for direct translation. However, the score for the Ko $\rightarrow$ Zh pair improved; thus, we can conclude that the pivot translation can achieve quality close to that of direct translation.

The zero-shot translations, on the other hand, showed very low scores for the unlearned language pairs. The scores were higher for the learned language pairs (Ja $\leftrightarrow$ X), although they were lower than those for direct translation. This means that the multilingual (unified) model was learned reasonably well; however, further study is required because zero-shot translation has only been researched for a few years.

		Target Language			
		Ja	En	Zh	Ko
Source Language	Ja	–	27.31	35.60	52.81
	En	30.84	–	22.21	26.27
	Zh	42.33	24.14	–	34.85
	Ko	57.92	24.85	30.66	–

(a) Direct Translation

		Target Language			
		Ja	En	Zh	Ko
Source Language	Ja	–	(27.31)	(35.60)	(52.81)
	En	(30.84)	–	20.82	25.49
	Zh	(42.33)	23.85	–	33.87
	Ko	(57.92)	24.55	30.82	–

(b) Pivot Translation
The brackets denote scores of the direct translation.

		Target Language			
		Ja	En	Zh	Ko
Source Language	Ja	–	26.32	34.17	51.73
	En	29.49	–	2.90	5.11
	Zh	40.71	8.87	–	10.00
	Ko	56.66	10.63	6.44	–

(c) Zero-shot Translation (En, Zh, and Ko)

Table 5: Comparison among Direct, Pivot, and Zero-shot Translation

6. Future Directions

First, we intend to further increase the size of the corpus to improve translation quality. Because the amount of the current corpus spreads various range among languages, we especially complement Asian languages, which is a feature of our corpus.

One current problem is context-dependent translation. We consider that there are two types of context-dependent translation.

- One type of context-dependent translation depends on external knowledge, such as domain knowledge. Such translations use special words and expressions depending on the domain. For example, there are two English translations of the following Japanese sentence in the disaster prevention domain.

  Ja1 *Youyaku yure-ga osamari-mashita.*
  finally shaking-SUBJ finish-POLITE
  En1-1 *The shaking finally has calmed down.*
  En1-2 *The earthquake is now stopped.*

  En1-1 is a literal translation of **Ja1**. **En1-2** can only be used when shaking is caused by an earthquake, and it is a particular translation within the disaster prevention domain.

- The other type of context-dependent translation depends on previous utterances. Here, the information of a source sentence is complemented or eliminated in a translation. For example, one translation of **Ja2** is **En2-2** in the GCP Corpus.

  Ja2 *Moyori eki-wa doko-desuka*?
  nearest station-TOP where-POLITE
  En2-1 *Where is the nearest station?*
  En2-2 *What is the station closest to the park?*

  "*The park*" in **En2-2** are extra words of **Ja-2**. However, the speaker also spoke "*I've never heard Tetsugakudo Park.*" in the preceding utterance, so it is correct in this context.

Human translators tend to translate literally between languages of the same family, such as English and French. In contrast, with language pairs for which it is difficult to make literal translations, such as English and Japanese, professional translators elaborately generate context-dependent translations to make the translations natural and the meaning of the dialogue identical.

Most MTs assume that the translation unit is a sentence. Therefore, it is harmful if a training corpus contains bilingual sentences with extra or missing words. The first type of context-dependent translation, which refers to external knowledge, is being solved by domain adaptation techniques. The second type, which depends on previous utterances, has not been solved using a sentence as the translation unit.

In the GCP Corpus, we attempt to apply the following counterapproach to maintain fluency and reduce extra/missing words. [4] First, utterances in a dialogue are shuffled and human translators translate them. This process breaks context in a dialogue; however, translators can refer to external knowledge if they read the entire dialogue. Then, the utterance order is restored, and the translations are checked to maintain the fluency of the dialogue.

7. Conclusion

In this paper, we have introduced the GCP Corpus, which is being developed as part of Global Communication Plan. The GCP Corpus is a multilingual sentence-aligned corpus that is being developed to help realize speech translation. The GCP Corpus has the following characteristics.

- It covers ten languages: Japanese, English, Chinese, Korean, Thai, Vietnamese, Indonesian, Myanmar, Spanish, and French. Notably, it includes Asian languages.
- It supports four target domains, i.e., medical care, disaster prevention, shopping, and tourism.
- It primarily consists of pseudo-dialogues between foreign visitors and local Japanese people.

Here, we have focused on the first of these characteristics and have investigated the corpus by evaluating the quality of different MT systems. In addition, we have compared direct, pivot, and zero-shot translations by taking advantage of the fact that the GCP Corpus is a parallel corpus.

[4] We started this process in the middle of the construction.

As we work toward the 2020 Tokyo Olympic and Paralympic Games, we will increase the size of the GCP Corpus, use it to develop speech translation systems, and distribute these systems widely [5].

8. Acknowledgments

This work was supported by "Promotion of Global Communications Plan — Research and Development and Social Demonstration of Multilingual Speech Translation Technology," a program of the Ministry of Internal Affairs and Communications, Japan.

9. Bibliographical References

Cohn, T. and Lapata, M. (2007). Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 728–735, Prague, Czech Republic, June.

Imamura, K. and Sumita, E. (2016). Multi-domain adaptation for statistical machine translation based on feature augmentation. In *Proceedings of the 12th Conference of the Association for Machine Translation in the Americas (AMTA-2016): Volume 1' MT Researchers' Track*, pages 79–92, Austin, Texas, USA, October.

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. (2016). Google's multilingual neural machine translation system: Enabling zero-shot translation. ArXiv e-prints. 1611.04558.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada, July.

Morishita, M., Suzuki, J., and Nagata, M. (2017). NTT neural machine translation systems at WAT 2017. In *Proc. of WAT2017*, pages 89–94.

Oda, Y., Sudoh, K., Nakamura, S., Utiyama, M., and Sumita, E. (2017). A simple and strong baseline: NAIST-NICT neural machine translation system for WAT2017 English-Japanese translation task. In *Proc. of WAT2017*, pages 135–139.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, Pennsylvania, USA.

Sennrich, R., Haddow, B., and Birch, A. (2016). 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.

Utiyama, M. and Isahara, H. (2007). A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 484–491, Rochester, New York, April.

10. Language Resource References

Higashinaka, R., Funakoshi, K., Kobayashi, Y., and Inaba, M. (2016). The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, May.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X*, pages 79–86, Phuket, Thailand.

[5] We are providing the VoiceTra system.
`http://voicetra.nict.go.jp/`