BehanceMT: A Machine Translation Corpus for Livestreaming Video Transcripts

Minh Van Nguyen[1], **Franck Dernoncourt**[2], **and Thien Huu Nguyen**[1]
[1] Dept. of Computer and Information Science, University of Oregon, Eugene, OR, USA
[2] Adobe Research, Seattle, WA, USA
{minhnv,thien}@cs.uoregon.edu, dernonco@adobe.com

Abstract

Machine translation (MT) is an important task in natural language processing, which aims to translate a sentence in a source language to another sentence with the same/similar semantics in a target language. Despite the huge effort on building MT systems for different language pairs, most previous work focuses on formal-language settings, where text to be translated come from written sources such as books and news articles. As a result, such MT systems could fail to translate livestreaming video transcripts, where text is often shorter and might be grammatically incorrect. To overcome this issue, we introduce a novel MT corpus - BehanceMT for livestreaming video transcript translation. Our corpus contains parallel transcripts for 3 language pairs, where English is the source language and Spanish, Chinese, and Arabic are the target languages. Experimental results show that finetuning a pretrained MT model on BehanceMT significantly improves the performance of the model in translating video transcripts across 3 language pairs. In addition, the finetuned MT model outperforms GoogleTranslate in 2 out of 3 language pairs, further demonstrating the usefulness of our proposed dataset for video transcript translation. BehanceMT will be publicly released upon the acceptance of the paper.

1 Introduction

Machine Translation (MT) is an important and challenging task in natural language processing. Early work solved the task via statistical models (Al-Onaizan et al., 1999; Och et al., 2004; Lopez, 2008; Koehn, 2009). Recent work has made significant improvement via deep learning models (Luong et al., 2015; Vaswani et al., 2017; Devlin et al., 2019; Yang et al., 2019; Lewis et al., 2020) that formalize MT as a text generation task, where an encoder is used to consume input text in a source language and a decoder is employed to generate the input's translation in a target language. In addition to the advance in model design, another factor contributing to the success of deep learning models is the creation of enormous MT corpora for model training such as WMT corpora (Bojar et al., 2014, 2016), OPUS corpus (Tiedemann, 2012) and IWSLT corpus (Cettolo et al., 2015). However, these corpora often contain formal-language texts such as books and news articles. This could lead to poor performance of the MT models, which are pretrained on such corpora, on informal-language text such as video transcripts. This is unfortunate as video transcripts are being generated at growing rate in international online video platforms such as Youtube [1], Dailymotion [2], and Behance [3]. Video transcript translation is thus important to improve access to the platforms' content for users who speak different languages.

In this work, we aim to address this issue by introducing a novel MT corpus - BehanceMT for video transcript translation (VTT). BehanceMT contains transcripts collected from the Behance platform and translations obtained by human annotators for 3 language pairs, where English is the source language and Spanish, Chinese, and Arabic are the target languages. An MT system pretrained on formal-language corpora can then be finetuned on BehanceMT to improve its performance for VTT. To demonstrate this idea, we employ OpusMT (Tiedemann and Thottingal, 2020), which is a popular MT system pretrained on OPUS corpora. For each language pair, we finetune the pretrained OpusMT on the BehanceMT training data and evaluate the model (called OpusMT+) on the test data. Experimental results show that OpusMT+ consistently outperforms OpusMT in all settings across the three language pairs for VTT. In addition, we compare OpusMT+ with Google-

[1] https://www.youtube.com/
[2] https://www.dailymotion.com/
[3] https://www.behance.net/

Proceedings of the 29th International Conference on Computational Linguistic, pages 30–33
October 12–17, 2022.

Translate [4]. The significant improvement obtained by OpusMT+ over GoogleTranslate in English → Chinese and English → Spanish further demonstrates the usefulness of our proposed MT corpus. To facilitate future work for VTT, we will publicly release the BehanceMT corpus.

2 Related Work

Previous work has created different corpora for MT, such as WMT corpora (Bojar et al., 2014, 2016), OPUS corpus (Tiedemann, 2012) and IWSLT corpus (Cettolo et al., 2015). However, most of these corpora focus on formal-language settings. To the best of our knowledge, (Cettolo et al., 2015), which involves parallel TED talks, is the closest work to ours. However, TED talks are mostly presented in formal language. By contrast, BehanceMT is created based on transcripts of livestreaming videos, which are more informal.

3 Data

In this section, we present how we collect, preprocess, and annotate video transcripts to create the BehanceMT corpus.

3.1 Data Collection

Video transcripts in the BehanceMT corpus are collected from livestreaming videos on Behance, a platform for livestreaming tutorial videos on creative works such as digital drawing, graphic design, and photo/video editing. Each video transcript contains multiple sentences produced by the Microsoft Automatic Speech Recognition (ASR) system (Xiong et al., 2018). To achieve a diverse corpus given a fixed annotation budget, we randomly select 99 video transcripts and retain at most 50 first sentences with an average length of 10 words for each transcript. The resulting transcripts are finally used to perform data annotation.

3.2 Data Annotation

To translate the video transcripts, we hire crowdsourcing workers on Upwork [5], who are native speakers of the target languages and proficient in English. Particularly, two crowd-sourcing workers are hired for translating video transcripts to Spanish, two crowd-sourcing workers are employed for translating video transcripts to Arabic, and one crowd-sourcing worker is hired for translating the video transcripts to Chinese. The workers are paid approximately $\$0.4$ for translating a sentence on average. Each worker performs the translation task by writing a translation for each sentence in an excel sheet containing their assigned video transcripts. To facilitate their annotation process, we also provide the video titles for each transcript so that the annotators can look up and watch the original videos if necessary.

Finally, we randomly split the translated video transcripts into train/dev/test parts with a ratio of $80/10/10$ for model development. The statistics for the resulting BehanceMT corpus is shown in Table 1.

Data	#transcripts	#sentences	#tokens
Train	78	3,787	40,024
Dev	11	530	5,007
Test	10	449	4,617

Table 1: Statistics for English data in BehanceMT corpus. Data for the target languages (Spanish, Arabic, and Chinese) contains the translations for each sentence in the English data.

4 Model

We employ OpusMT (Tiedemann and Thottingal, 2020) as the main model to conduct experiments on the proposed BehanceMT corpus. OpusMT uses the Marian-NMT architecture (Junczys-Dowmunt et al., 2018) and is pretrained on OPUS corpus (Tiedemann, 2012) to perform the translation task for different language pairs. For each of the three language pairs (i.e., English → Spanish, English → Arabic, English → Chinese), we further finetune the pretrained bilingual OpusMT model on the corresponding training data in BehanceMT. We denote the finetuned OpusMT model as OpusMT+.

5 Experiments

5.1 Model Training and Hyper-parameters

To implement the models, we use Pytorch 1.12.1 and Huggingface Transformers 4.21.1. The pretrained OpusMT models *"opus-mt-en-es"*, *"opus-mt-en-ar"*, and *"opus-mt-en-zh"* are obtained respectively for English → Spanish, English → Arabic, and English → Chinese settings from the official model hub [6]. To finetune the models on BehanceMT data, we employ Adam optimizer (Kingma and Ba, 2015) to train the model for 50

[4] https://translate.google.com/
[5] https://www.upwork.com/
[6] https://huggingface.co/Helsinki-NLP

epochs with a batch size of 16, a learning rate of $1e-6$, and a weight decay of 0.01.

Models	Spanish	Chinese	Arabic
OpusMT	35.0	5.3	25.2
OpusMT+	37.5	13.7	33.4
GoogleTranslate	34.9	3.1	43.2

Table 2: Model performance (BLEU score) comparison on BehanceMT test sets for the three target languages.

5.2 Performance Comparison

Table 2 presents performance comparison between OpusMT, OpusMT+, and GoogleTranslate across the three language pairs on test sets of our proposed BehanceMT corpus. First, we can see that OpusMT and GoogleTranslate perform poorly in most settings. This suggests that VTT is challenging task and more research effort is necessary to improve the performance for this area. Second, OpusMT+ significantly outperforms OpusMT in all settings, showing the benefit of finetuning OpusMT on video transcript data for improving model performance for VTT. This is further confirmed as OpusMT+ obtains significant improvement compared to the state-of-the-art commercial translation engine GoogleTranslate in two out of the three translation settings.

6 Conclusion

In this work, we present a novel corpus - BehanceMT for video transcript translation (VTT). Behance contains parallel video transcripts for three language pairs, where English is the source language and Spanish, Arabic, and Chinese are the target languages. Our experiments with strong baselines on BehanceMT show that the proposed corpus is challenging and useful for VTT across the three language pairs.

References

Yaser Al-Onaizan, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, Dan Melamed, Franz-Josef Och, David Purdy, Noah A Smith, and David Yarowsky. 1999. Statistical machine translation. In *Final Report, JHU Summer Workshop*, volume 30.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Ales Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2015. The IWSLT 2015 evaluation campaign. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 2–14, Da Nang, Vietnam.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Philipp Koehn. 2009. *Statistical machine translation*. Cambridge University Press.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Adam Lopez. 2008. Statistical machine translation. *ACM Computing Surveys (CSUR)*, 40(3):1–49.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alexander Fraser, Shankar Kumar, Libin Shen, David A Smith, Katherine Eng, et al. 2004. A smorgasbord of features for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 161–168.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann and Santhosh Thottingal. 2020. Opus-mt–building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wayne Xiong, Lingfeng Wu, Fil Alleva, Jasha Droppo, Xuedong Huang, and Andreas Stolcke. 2018. The microsoft 2017 conversational speech recognition system. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5934–5938. IEEE.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.