NICT's Submission to the WAT 2022 Structured Document Translation Task

Raj Dabre
National Institute of Information and Communications Technology (NICT),
Kyoto, Japan
raj.dabre@nict.go.jp

Abstract

We present our submission to the structured document translation task organized by WAT 2022. In structured document translation, the key challenge is the handling of inline tags, which annotate text. Specifically, the text that is annotated by tags, should be translated in such a way that in the translation should contain the tags annotating the translation. This challenge is further compounded by the lack of training data containing sentence pairs with inline XML tag annotated content. However, to our surprise, we find that existing multilingual NMT systems are able to handle the translation of text annotated with XML tags without any explicit training on data containing said tags. Specifically, massively multilingual translation models like M2M-100 perform well despite not being explicitly trained to handle structured content. This direct translation approach is often either as good as if not better than the traditional approach of "remove tag, translate and re-inject tag" also known as the "detag-and-project" approach.

1 Introduction

Neural machine translation (Bahdanau et al., 2015) using transformers (Vaswani et al., 2017) is gradually beginning to reach a saturation point in terms of translation quality for major languages like English, French, Japanese, Chinese (Fan et al., 2021). Most existing work focus on the translation of plain text, where the sentence is translated individually or by considering its context via a document level translation approach (Miculicich et al., 2018). However, this does not directly address an important real life application: "web page translation". Web pages are structured documents containing formatted or annotated text, where the annotation is done via inline tags or XML tags. When translating web pages, care must be taken to translate not only the text but also the XML tags. For example, *This is a <b>sentence</b>.* is an example of a sentence in a structured document. Its translation in Spanish should be: *Esta es una <b>frase</b>.* where the <b> and </b> tags appropriately enclose the translation of the word *sentence* which is *frase*. The structured document translation task[1] in WAT 2022 aims at evaluating approaches for the translation of text with XML tags or inline tags. For a detailed overview of the task, kindly refer to the overview paper (Nakazawa et al., 2022).

Since NMT models are sensitive to what they are trained on, it is natural to assume that they should be exposed to examples of how to handle XML tags. Unfortunately, there is a scarcity of training data containing XML tags to train NMT models to handle structured content. Hashimoto et al. (2019) provide training data for 7 languages, but this is not possible for all languages. Therefore, the most viable solution would be the "remove tag, translate and re-inject tag" approach also known as the detag-and-project approach (Zenkel et al., 2021) shortened to DnP. The main problem with DnP is that it needs high quality word alignments and heuristic algorithms when reinserting the tags into the translation. Therefore, poor translations, poor alignments and heuristics lead to compounding errors which can negatively affect the injection process leading to poor transfer of structure.

In WAT 2022, we participated under the team name "NICT-5" where we applied the DnP approach to the structured document translation task for English to Japanese/Chinese/Korean as well as Japanese/Chinese/Korean to English translation. Given the large availability of pre-trained translation models, we decided to use the M2M-100 model. In order to compare against the DnP approach, we translated sentences containing XML tags using this model and to our surprise, this approach was able to outperform the DnP approach in some instances. Our analyses reveal that the

[1] https://lotus.kuee.kyoto-u.ac.jp/WAT/NICT-SAP-Task/index.2022.struc.html

DnP approach is better at transferring the XML tag structure but gives poor automatic evaluation scores in some cases as it fails to handle cases such as non-closing tags and the tag injection for words and phrases for which word alignment fails.

2 Related Work

Hashimoto et al. (2019) present a dataset from the IT domain, same as the domain of the evaluation set used in the task, that features XML markup, and corresponding results using a constrained beam search approach for decoding. They create and use training data with XML tags but we do not and instead opt to use direct translation and the detag-and-project (DnP) approaches Hanneman and Dinu (2020). The methods for tag transfer in Zenkel et al. (2021) are relevant, although their focus is on inserting tags into a fixed human translation. Although the task evaluation sets contain complete documents which can allow for context-sensitive translation, such as in Miculicich et al. (2018), and in-context evaluation (Läubli et al., 2018, amongst others), we do not focus on these aspects in our submission.

In terms of methods, according to our knowledge, we are the first to report results on tag transfer using a massively multilingual translation model like M2M-100 (Tang et al., 2021; Fan et al., 2021) which surprisingly lead to reasonable automatic evaluation scores. We also submit results for the DnP approach, but find that it does not always outperform the direct translation approach. For evaluation, WAT uses the XML-BLEU metric in accordance with Hashimoto et al. (2019) but we additionally report on the XML tag structure accuracy to better understand the limitations of the approaches we used.

3 Approaches

We use the direct translation (DiT) and the detag-and-project (DnP) approaches for our submissions.

a. Direct translation: In this approach, we directly translate the sentences with XML content in them.

b. Detag-and-project: In this approach, we use the following steps:

1. Remove the XML tags from the sentence and make a list of words and phrases which are wrapped with XML tags. In case of non-closing tags, we do not handle them.
2. Translate the plain sentences.
3. Use a word aligner to align the words between the plain sentence and its translation.
4. For each sentence, for each word or phrase obtained in step 1, get its aligned target word and phrase and wrap it with the applicable tag.

Note that the following considerations are to be made:

- For translation, the NMT model's tokenizer can handle subword segmentation.
- For word alignment, tokenizers should be used for unsegmented languages prior to alignment.
- To infer phrase alignment, we use the inside-outside algorithm from Zenkel et al. (2021) who also used an alternative approach called the min-max algorithm, but we do not use it in our submissions as we found the former to be slightly better.
- When translating content wrapped hierarchically in XML tags, the innermost tags are dealt with first.

4 Experiments

4.1 Dataset

We only use the official development and test sets (located here) provided by the organizers. We focus on translation to and from English and Japanese/Korean/Chinese. We do not consider traditional Chinese due to lack of reliable models and word aligners.

4.2 Implementation

We implement the inside-outside approach in Python along with other pre-processing scripts. For word alignment we use awesome-align (Dou and Neubig, 2021).[2] as we do not have reliable training data for word alignment. awesome-align uses mBERT[3] and is known to work well even without using fine-tuning to improve alignment quality. For tokenization prior to word alignment, we use mecab for Japanese[4] and Korean[5] and Stanford

[2] https://github.com/neulab/awesome-align

[3] https://github.com/google-research/bert/blob/master/multilingual.md

[4] https://taku910.github.io/mecab/

[5] https://github.com/SamuraiT/mecab-python3

XML-BLEU

Approach	**en→ja**	**ja→en**	**en→ko**	**ko→en**	**en→zh**	**zh→en**
DnP	36.84	25.02	22.81	23.80	32.34	28.50
DiT	36.40	18.76	**28.99**	**24.35**	**32.38**	29.06
Organizer	**40.27**	**28.20**	21.87	10.80	28.03	**29.14**

XML structure transfer accuracy (%)

Approach	**en→ja**	**ja→en**	**en→ko**	**ko→en**	**en→zh**	**zh→en**
DnP	**84.38**	**85.35**	**86.93**	**80.24**	**85.40**	**84.16**
DiT	81.66	23.51	82.34	77.85	83.53	81.26

Table 1: XML-BLEU and XML structure transfer accuracy scores for our submissions and the organizer submission. Best scores are in bold.

segmenter for Chinese[6].

4.3 Models Used

We use the M2M-100 (or M2M) 1.2 billion parameter model[7] (Fan et al., 2021) which supports 100 languages. To our knowledge, M2M was not trained to handle XML tags in sentences. We use beam search with beam size 4 and length penalty of 1.0.

4.4 Evaluation

WAT uses the XML-BLEU metric proposed by Hashimoto et al. (2019) using a modified version[8] of the publicly available repository. The modification was done to handle the XML tags specific to the evaluation sets. We use this modified code for our analyses as well. Specifically, we calculate the XML structure transfer accuracy, which indicates the number of sentences whose XML structures have been transferred into the translation. This only concerns the structure and not the content wrapped in the XML tags. There are 590 sentences in the test set with XML tags in them and the accuracy indicates the percentage of sentences with proper structure transfer.

5 Results

We present in Table 1 the XML-BLEU scores and the XML structure transfer accuracy for our submissions. In the last row, we give the organizer scores. According to their description, they seem to use am mBART model for direct translation (DiT). The results show that except for Japanese↔English translation and Chinese→English translation, our submissions are better than the organizer's submissions. The organizer scores for Japanese↔English translation are vastly better than ours for this direction, and this may be due to the ability of the mBART model they used to translate to/from Japanese better than M2M-100. Indeed, for Chinese→English translation, the gap between our best and organizers is 0.08 XML-BLEU which is negligible. For the remaining directions, our submissions are substantially better by at least 4.35 XML-BLEU.

Comparing the DnP and DiT rows for our submissions, it can be seen that except for Japanese↔English translation, DiT is slightly if not substantially better than DnP. This is quite surprising since the M2M model was never explicitly trained to handle XML tags. It is possible that the model treats the tags as rare or unknown English tokens which are usually copied as is in Japanese, Chinese and Korean translation. We leave this investigation for the future.

With regard to the XML structure transfer accuracy, it is interesting that although the XML-BLEU is higher for DnP, the structure transfer accuracy is lower. Upon some manual investigation we found the following:

- DnP is good at transferring structure but is bad at transferring it in the right place. This is due to the difficulty in aligning phrases which is affected by language divergence and word alignment quality. Using a high quality word aligner should help resolve this partially.

- Whenever DnP is unable to align words or phrases, the entire example wont count towards the structure match accuracy. This happens in case of non-closing tags which we do

[6] https://nlp.stanford.edu/software/stanford-segmenter-4.2.0.zip

[7] https://huggingface.co/facebook/m2m100_1.2B

[8] https://github.com/prajdabre/localization-xml-mt

not transfer as they do not wrap any word or phrase making it hard, if not impossible, to determine its position in the translation. However, this problem does not occur for DiP.

- DiP often hallucinates tags or discards them. Since our NMT model was not explicitly trained to handle tags, this makes sense. Some constrained decoding would be helpful here.

Overall, the DnP approach needs a lot of investment but the returns are not equivalent. Future work should focus more on the DiT approach which is end-to-end and hence more attractive.

6 Conclusion

In this paper, we describe our submissions as team "NICT-5" to the structured document translation task in WAT 2022. We used the direct translation and the detag-and-project approaches and to our surprise found that the direct translation approach outperforms detag-and-project approach slightly or substantially depending on the language pair. Our analyses reveal that the former approach has poorer tag structure transfer accuracy, but still is better than the latter approach, due to (a.) the latter's inability to handle the transfer of tags for content that can't be aligned with its translation and (b.) the latter's sensitivity to poor alignment. Rather than working to improve the detag-and-project approach, we plan to focus more on the direct translation approach with constrained generation and some additional training to handle structured content more effectively.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Greg Hanneman and Georgiana Dinu. 2020. How should markup tags be translated? In *Proceedings of the Fifth Conference on Machine Translation*, pages 1160–1173, Online. Association for Computational Linguistics.

Kazuma Hashimoto, Raffaella Buschiazzo, James Bradbury, Teresa Marshall, Richard Socher, and Caiming Xiong. 2019. A high-quality multilingual dataset for structured documentation translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 116–127, Florence, Italy. Association for Computational Linguistics.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Toshiaki Nakazawa, Hideya Mino, Isao Goto, Raj Dabre, Shohei Higashiyama, Shantipriya Parida, Anoop Kunchukuttan, Makoto Morishita, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. 2022. Overview of the 9th workshop on Asian translation. In *Proceedings of the 9th Workshop on Asian Translation (WAT2022)*, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Thomas Zenkel, Joern Wuebker, and John DeNero. 2021. Automatic bilingual markup transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3524–3533, Punta Cana, Dominican Republic. Association for Computational Linguistics.