A Multilingual Multiway Evaluation Data Set for Structured Document Translation of Asian Languages

Bianka Buschbeck[†*] **Raj Dabre**[‡*] **Miriam Exel**[†] **Matthias Huck**[†]

Patrick Huy[†] **Raphael Rubino**[‡] **Hideki Tanaka**[‡]

[†]SAP SE
Dietmar-Hopp-Allee 16
69190 Walldorf
Germany
firstname.lastname@sap.com

[‡]NICT
3-5 Hikaridai
Seika-cho, Soraku-gun, Kyoto, 619-0289
Japan
firstname.lastname@nict.go.jp

Abstract

Translation of structured content is an important application of machine translation, but the scarcity of evaluation data sets, especially for Asian languages, limits progress. In this paper we present a novel multilingual multiway evaluation data set for the translation of *structured documents* of the Asian languages Japanese, Korean and Chinese. We describe the data set, its creation process and important characteristics, followed by establishing and evaluating baselines using the direct translation as well as detag-project approaches. Our data set is well suited for multilingual evaluation, and it contains richer annotation tag sets than existing data sets. Our results show that massively multilingual translation models like M2M-100 and mBART-50 perform surprisingly well despite not being explicitly trained to handle structured content. The data set described in this paper and used in our experiments is released publicly.

1 Introduction

A common use case of machine translation (MT) is the translation of structured or formatted documents, such as web pages. The key challenge is to properly transfer markup tags *within* the translatable content (e.g. bold) from the source to the target language during the translation process. A markup example is shown in Figure 1. Although there are various data sets for sentence- and document-level machine translation, apart from Hashimoto et al. (2019) and Hanneman and Dinu (2020) we are not aware of any other data sets for evaluating the translation quality of markup annotated sentences. This paper introduces a data set that reflects all those aspects to facilitate and foster research that goes beyond the translation of plain text in isolation.

en	Click <uicontrol>Prepayment</uicontrol>.
ja	<uicontrol>前払</uicontrol>をクリックします。

Figure 1: Example with inline markup (in gray).

In this paper, we describe the second release of the *software documentation data set for machine translation*, a high-quality multilingual evaluation data set for machine translation in the IT domain.[1] It has been released by SAP[2], a large enterprise software company. The contents originate from the *SAP Help Portal*[3] that contains documentation and learning materials for SAP products. With this release of the data set, we publish development and test data for MT purposes in the form of *complete structured documents* that include segment-internal (inline) markup, in a rich XML-based localization format as well as transformations that make it readily usable in many standard machine translation workflows. It consists of 385 documents that contain about 4,000 translatable segments and their translations. With the second release, we focus on the following major Asian languages: Japanese (ja), Korean (ko), and Chinese (zh). Translations have been produced from the same English (en) source, thus the data is multiway parallel. The multiway document-level nature of this data set enables not only evaluation of multilingual models but also document-level translation approaches (if needed) when translating structured content.

Additionally, in this paper, we establish base-

[*] Equal contribution. Ordered by last name.

[1] The *software documentation data set for machine translation* is available under the Creative Commons license Attribution-Non Commercial 4.0 International CC BY-NC 4.0). The second release can be downloaded from https://github.com/SAP/software-documentation-data-set-for-machine-translation/releases/tag/v2.1.

[2] https://www.sap.com/

[3] https://help.sap.com/

Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, pages 237–245
November 20–23, 2022.

lines for the released data set for individual segment translation, where we utilize massively multilingual models such as M2M-100 (Fan et al., 2021) and mBART-50 (Tang et al., 2021), making use of *out-of-the-box* publicly available checkpoints already trained in a many-to-many translation fashion with no additional fine-tuning on our end. We show that these models can be used for directly translating structured content despite not being explicitly trained to do so. We observe that the quality of the direct translation approach, where the source text is composed of both lexical and markup content, is comparable to the traditional detag-project approach. We then report translation results according to several metrics targeting not only the translation quality but also tag placement accuracy, allowing us to understand the difficulty of translating structured content into Asian languages.

2 Related Work

Only recently awareness has increased that real world content often resides in structured and formatted documents such as HTML pages and Microsoft Office formats, and that the transfer of inline markup tags is a challenge for neural machine translation; correspondingly, little work has been published. Hashimoto et al. (2019) present a data set from the IT domain that features inline markup, and corresponding MT results using a constrained beam search approach for decoding. Furthermore, Hanneman and Dinu (2020) compare different data augmentation methods with a detag-project approach, and evaluate on data from legal documents from the European Union. The methods for tag transfer in Zenkel et al. (2021) are also related, even though they focus on inserting the tags into a fixed human translation.

In contrast to the previously mentioned available data sets, with the *software documentation data set for machine translation*, we publish complete documents of high translation quality, thus allowing for context-sensitive translation, such as in Miculicich et al. (2018) for example, and in-context evaluation as it has been shown to be vital for accurate evaluation assessments (Läubli et al., 2018, amongst others). Furthermore, our data set is multiway multilingual, focuses on Asian languages and adds lower resource Asian languages to the picture to enable a more comprehensive evaluation of different methods. While Hashimoto et al. (2019) enables evaluation of 14 translation directions to or from English, they do not support non-English translation directions as their evaluation data is not n-way parallel. In contrast, the second release of the *software documentation data set* is n-way parallel enabling 6 translation directions to and from English as well as 6 translation directions between the Asian languages leading to a total of 12 directions.

The first release of the *software documentation data set for machine translation* is described in Buschbeck and Exel (2020). While it also contains complete documents with rich metadata on the segment level and is therefore well suited to evaluate contextual approaches to MT, it does not feature complete hierarchical document structure. Its focus is low-resource language pairs that are typically under-represented in MT research, namely English to Hindi, Indonesian, Malay and Thai.

In terms of methods, according to our knowledge, we are the first to report results on tag transfer using pre-trained massively multilingual translation models mBART-50 and M2M-100 (Tang et al., 2021; Fan et al., 2021). We also compare with the detag-project approach, but leave the exploration of other methods on this data set for future work.

3 The Structured Documents Data Set

We describe the second release of the *software documentation data set for machine translation*, our data set for structured document translation of Asian languages.

3.1 Data Set Sources and Selection

The contents of the data set originate from the public online documentation of SAP, a large software company, featuring product documentation, user assistance and learning materials. The individual pages (or documents) are highly structured. They are authored in DITA, an XML-based open standard often used for technical documentation.[4] The original documents are in English and translations are performed into Japanese, Korean and Chinese (amongst others) by specialized professional translators. Translations are validated in a subsequent review process to guarantee an excellent quality, including coherent domain-specific terminology, before the final target texts are published. Throughout this process, standard computer-assisted translation tools are used. The localization workflow is based

[4] https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=dita

on XLIFF, an XML-based format for storing bitext which was created to standardize the way localizable data is passed between localization tools.[5] The original DITA document structure including inline markup is preserved throughout the process. For more background information on the data, consult Buschbeck and Exel (2020).

Documents for development and test data are selected from a large set of original DITA documents that have recently been translated, with the same English source for all target languages. To create an interesting and relevant data set, we calculate a set of indicators per document, and then select those documents that score best. In order to minimize segment redundancy within the data set (ratio of all source-target pairs to unique source-target pairs) while selecting complete documents, we follow the criteria introduced in Buschbeck and Exel (2020). Besides document length and average segment length, they consider the redundancy within documents as well as between documents. In addition, we also take the number of inline markup tags into account.

3.2 Format and Tooling

We provide the data in XLIFF. Each XLIFF file of the data set represents one original DITA document with its translation into one of the target languages. Appendix A.1 provides more details, including an example. Our XLIFF files contain the *full* original document structure and are therefore very rich in information. However, some applications or evaluation scenarios might only want to consider specific parts of the structure. Therefore, we also provide the data in a format that is convenient for MT research: one translatable (source or target) segment per line with inline markup being represented as raw DITA tags, similar to the format in Hashimoto et al. (2019), an example of which is in Table 1. This representation is obtained from XLIFF with an XSL transformation. Other transformations for which we provide XSL stylesheets are described in Appendix A.2.

3.3 Data Set Statistics and Characteristics

Table 1 displays the main characteristics of the data set such as number of documents, translatable segments, segments containing inline elements, number of words and amount of redundancy. As the data sets are composed of whole documents, some segment duplicates are unavoidable, despite a data selection method that strives for a low intersection of documents (see Section 3.1). In the data at hand, we were not able to avoid the same headings that occur across documents. For example, the heading *Use* occurs 96 times and *Definition* 49 times in the test set. The rest of the segments are mostly unique. Additional statistics can be found in Appendix A.3.

The DITA inline elements of the data set are provided in Table 2. Most of them consist of an opening and a closing tag, such as *<uicontrol>...</uicontrol>*, others are self-closing, e.g. *<xref keyref=... />*. There are a total of 27 different types of inline elements that serve different purposes: many are formatting and style markers,

	dev	test
Number of documents	195	190
Number of parallel transl. segments	2,011	2,002
↪ containing inline elements	520	590
Number of source words	24,490	24,244
Segment redundancy	1.09	1.08

Table 1: Characteristics of development and test sets for the English source of the second release of the *software documentation data set for machine translation*.

Type	dev	test
alt	2	2
cite	27	8
codeph	1	7
emphasis	37	55
field	1	3
i	12	2
image	0	1
key	0	2
keys	0	1
keyword	1	0
menucascade	1	6
ph	1	0
pname	4	15
q	0	2
sap-icon-background-color	2	10
sap-icon-font	2	10
sap-icon-font-character	2	10
sap-icon-font-color	2	10
sap-icon-font-description	2	10
sap-icon-font-size	2	10
sap-note	2	0
sap-technical-name	25	41
systemoutput	8	1
uicontrol	569	647
uinolabel	3	9
userinput	13	16
xref	25	25

Table 2: Different types of inline elements present in development and test sets.

[5] http://docs.oasis-open.org/xliff/v1.2/os/xliff-core.html

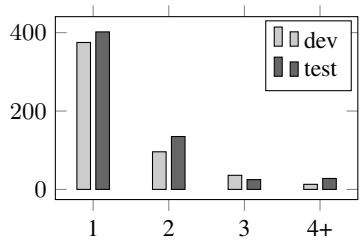

Figure 2: Distribution of inline elements per segment.

while others, such as *uicontrol*, *userinput* or *sap-technical-name*, are translation-relevant as they indicate if or how the annotated text should be translated. The most prevalent inline markup is *uicontrol*, used to mark up user interface controls, such as names of buttons, entry fields or menu items that require precise translation. Self-closing *xref* elements act as placeholders for text that is not accessible. Of the translatable segments in the dev and test data, 25.86% and 29.34%, respectively, contain at least one inline element. Figure 2 shows the number of sentences in dev and test sets containing one, two or more inline elements.

4 Baseline Experiments

We propose to evaluate the translation performance of out-of-the-box pretrained multilingual neural machine translation (NMT) systems for the English to Japanese, Korean and Chinese translation directions.[6] We focus on segment-level translation and propose to leave document-level approaches for future work.

4.1 NMT Models and Approaches

Publicly available multilingual translation models have shown to reach impressive results in terms of translation performance measured by popular automatic metrics. Due to the cost of training such models and in a bid to be eco-friendly, we use the M2M-100 (Fan et al., 2021) and the mBART-50 (Tang et al., 2021) many-to-many fine-tuned models which handle the translation directions of our data set. Both models are used from publicly available checkpoints to decode the data set with no additional fine-tuning on our end. Two hyperparameters are set for the decoder: a beam size of 4 and a length penalty of 1.0.

To handle mixed lexical and markup content, we consider two approaches:

Direct Translation (DT): We directly translate segments with markup using the NMT models.

Detag-project (DP): We first remove markup from the segments, translate segments, and insert the tags back into the translation using word alignments. We follow Zenkel et al. (2021) and use the inside-outside projection algorithm with alignments obtained from `awesome-align` (Dou and Neubig, 2021).[7]

4.2 Evaluation and Results

Previous work in structured document translation attempted to distill knowledge from widely used MT automatic metrics, such as BLEU (Papineni et al., 2002), by splitting content based on markup or measuring the accuracy of matching tags and attributes (Hashimoto et al., 2019; Hanneman and Dinu, 2020). In this work, we propose to maintain the commonly adopted evaluation approaches based on markup–lexis separation by allocating one metric per type of evaluation: **raw metrics**, computed on MT output and reference mixing text and markup, **lex metrics**, computed on MT output and reference stripped of markup, and **tag metrics**, computed on MT output and reference containing only markup. Note that the **raw** and **lex** metrics are similar to the *tagged* and *untagged* BLEU metrics, respectively, as proposed by Hanneman and Dinu (2020).

Overall comparison between two MT outputs can be conducted by comparing the raw metric scores, while the lex metric focuses on lexical tokens only and markup translation performance is measured by the tag metric. Table 3 reports the results obtained with SacreBLEU (Post, 2018) when computing BLEU following the three evaluation approaches listed above. Additional results using chrF (Popović, 2015) are presented in Table 4.[8]

Results obtained with the raw BLEU and chrF metrics show that both DT and DP approaches perform relatively well for two out of three transla-

[6] A total of 12 translation directions are available with the data released with this work.

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

[8] SacreBLEU signatures for raw and lex metrics:
Japanese BLEU:
nrefs:1|case:mixed|eff:no|tok:ja-mecab-0.996-IPA|smooth:exp|version:2.3.0
Korean BLEU:
nrefs:1|case:mixed|eff:no|tok:ko-mecab-0.996/ko-0.9.2-KO|smooth:exp|version:2.3.0
Chinese BLEU:
nrefs:1|case:mixed|eff:no|tok:zh|smooth:exp|version:2.3.0
tag BLEU:
nrefs:1|case:mixed|eff:no|tok:none|smooth:exp|version:2.3.0
chrF (all metrics):
nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.3.0

		raw			*lex*			*tag*		
		en→ja	en→ko	en→zh	en→ja	en→ko	en→zh	en→ja	en→ko	en→zh
M2M	DT	42.1	**34.6**	**49.2**	35.3	**27.1**	43.4	78.4	77.2	80.1
	DP	40.6	30.3	48.9	36.4	25.9	**44.7**	78.6	74.3	79.8
MBart	DT	**44.9**	28.5	44.3	37.2	18.9	37.8	**92.2**	**82.1**	**89.9**
	DP	41.5	26.8	44.3	**38.1**	19.6	39.1	78.8	79.4	79.4

Table 3: BLEU scores obtained with direct translation (DT) and detag-and-project (DP) using M2M and mBART models when evaluating mixed text and markup (*raw*), text only (*lex*) and markup only (*tag*).

		raw			*lex*			*tag*		
		en→ja	en→ko	en→zh	en→ja	en→ko	en→zh	en→ja	en→ko	en→zh
M2M	DT	53.2	**50.3**	**57.5**	40.2	34.2	37.5	91.4	**95.6**	93.3
	DP	54.0	47.7	57.3	42.3	**34.5**	**39.1**	94.7	94.3	94.5
MBart	DT	**57.4**	45.0	54.2	43.7	26.1	32.6	**96.3**	92.5	94.1
	DP	56.4	45.1	54.5	**45.6**	27.3	34.2	94.8	94.9	**94.8**

Table 4: chrF scores obtained with direct translation (DT) and detag-and-project (DP) using M2M and mBART models when evaluating mixed text and markup (*raw*), text only (*lex*) and markup only (*tag*).

ton directions tested, namely English-to-Japanese and English-to-Chinese. For the English-to-Korean translation direction, however, results for the lex BLEU metric indicate that M2M and MBart do not perform as well as for the two other translation directions, with MBart being outpeformed by M2M when Korean is the target.

The tag BLEU metric shows that MBart with the DT approach reaches the best results compared to the other approach and translation model. However, the tag chrF metric does not follow the same trend, which indicates that spacing within markup is better handled by MBart when translating tags in context (spaces are not taken into account with the chrF metric). The M2M model reaches the highest BLEU and chrF scores for Korean and Chinese target languages when lexical content is present (raw and lex metrics), while MBart reaches the highest scores when the target is Japanese.

Regardless of the metric (BLEU or chrF), DT exhibits better performance than DP in most cases, indicating that massively multilingual pre-trained MT systems can handle markup transfer without being explicitly trained on parallel data containing markup. DP, which involves tokenization, alignment and markup projection, involves imperfect heuristics (we have used inside-outside (Zenkel et al., 2021)). This makes DT without explicit training on markup data deserving of further exploration compared to DP. See Appendix B for additional results.

5 Conclusion

In this paper we have presented our multilingual multiway evaluation data set for structured document translation of three Asian languages, Japanese, Korean and Chinese – the second release of the *software documentation data set for machine translation*. Our data set contains rich annotation tag sets and is well suited for multilingual natural language processing tasks such as MT and its evaluation. We have established and evaluated MT baselines using two methods to handle inline markup, namely the direct translation and the detag-project approaches. Our results show that massively multilingual translation models like M2M-100 and mBART-50 perform surprisingly well despite not being explicitly trained to handle structured content. This previously unknown capability of MT models used in our experiments deserves further exploration, especially in combination with document-level translation approaches.

Acknowledgements

We would like to thank our SAP colleagues Ben Callard for extracting the DITA and XLIFF data from the corporate systems and Jens Scharnbacher for getting us started with XSLT. We are also thankful for the helpful feedback we got from the anonymous reviewers.

References

Bianka Buschbeck and Miriam Exel. 2020. A parallel evaluation data set of software documentation with document structure annotation. In *Proceedings of the 7th Workshop on Asian Translation*, pages 160–169, Suzhou, China. Association for Computational Linguistics.

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.

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

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. 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.

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.

A Data Set

We provide additional information on the second release of the *software documentation data set for machine translation*, such as the data format, available data transformations, and more data characteristics.

A.1 XLIFF Format

We provide the data in XLIFF (*.xlf*) Version 1.2. Each XLIFF file of the data set represents one original DITA document (*file* element) with its translation into one of the target languages. Within the *file* element, *trans-unit*s contain the localizable data: *source* elements store the source text, *seg-source* elements the (sentence) segmented source text and *target* elements the corresponding segments in the target language. Source and target segments are enclosed by *mrk* elements, and associated with each other via an ID (*mid* attribute). The full structure of the original DITA document is also represented in our XLIFF format. The DITA XML tags are enclosed by XLIFF inline elements (*ph*, *bpt*, *ept*). Much of the original DITA format can be restored by literally using the DITA tags masked by XLIFF inline elements. Whenever a *source* consists only of inline elements, the *translate* attribute of the enclosing *trans-unit* is set to *no*. When only parts of a translatable segment are not to be translated, this is represented as *<mrk mtype="protected">*. An example XLIFF document can be found in Figure 3. Information beyond the description here can be found in the Readme accompanying the data.

```
<?xml version="1.0" encoding="UTF-8"?>\\
<xliff xmlns="urn:oasis:names:tc:xliff:document:1.2" version="1.2">
<file original="dita" datatype="xml" source-language="en-US" target-language="ja
    -JP">
<body>
...
<trans-unit translate="no" id="feed189b-f66d-403d-84cd-068edc17edd1">
<source><ph id="18">&lt;/li&gt;</ph><ph id="19">&lt;li&gt;</ph></source>
</trans-unit>
<trans-unit id="32a07041-05f4-4e61-b4d3-1569b7b3509a">
<source>Click <bpt id="20">&lt;uicontrol&gt;</bpt>Prepayment<ept id="20">&lt;/
    uicontrol&gt;</ept>.
<seg-source><mrk mtype="seg" mid="7">Click <bpt id="20">&lt;uicontrol&gt;</bpt>
    Prepayment<ept id="20">&lt;/uicontrol&gt;</ept>.</mrk></seg-source>
<target><mrk mtype="seg" mid="7"><bpt id="20">&lt;uicontrol&gt;</bpt>前
    払<ept id="20">&lt;/uicontrol&gt;</ept>をクリックします。
    </mrk></target></trans-unit>
<trans-unit translate="no" id="ec9ffb5c-5516-4bb1-aa6a-bfafa5827bd0">
<source><ph id="21">&lt;/li&gt;</ph></source>
</trans-unit>
...
</body>
</file>
</xliff>
```

Figure 3: Excerpt of an XLIFF document (en-ja) of the data set.

A.2 Data Transformations

The released data in XLIFF contains the *full* document structure and is therefore very rich in information. However, some applications or evaluation scenarios might only want to consider specific parts of the structure. XLIFF documents can conveniently be transformed to different representations for different purposes using XSL stylesheets. For inspiration and convenience, we provide several stylesheets with the data that lead to the following transformed outputs:

(i) the structured document as a functional DITA file containing the source or target text and the original DITA tags;

(ii) one translatable (source or target) segment per line with inline markup being represented as DITA tags, similar to the format in Hashimoto et al. (2019);

(iii) one translatable (source or target) segment per line with inline markup being represented as XLIFF masking tags `x` and `g`, similar to the format in Hanneman and Dinu (2020);

(iv) one translatable (source or target) segment per line as plain text, without inline markup.

Examples for the transformations can be found in Figure 4. For convenience, we provide all source/-target documents concatenated after being transformed with method (iv) for standard machine translation evaluation, and with method (ii), as this format is relevant for current usage in machine translation research concerning tag transfer. The latter has been used in this work in Section 4.

(i)

```
</li><li>
Click <uicontrol>Prepayment</uicontrol>.
</li>
```

(ii)

```
Click <uicontrol>Prepayment</uicontrol>.
```

(iii)

```
Click <g  id="20">Prepayment</g>.
```

(iv)

```
Click Prepayment.
```

Figure 4: Data transformations (source) for the example in Figure 3.

The documents contain certain placeholders that reference textual content outside the respective document. In the plain-text data (iv), they have been replaced by `<locked-ref>` as just removing them would render the segments incomplete and ungrammatical.

		XML BLEU			*BLEU*			*Markup Matching %*		
		en→ja	en→ko	en→zh	en→ja	en→ko	en→zh	en→ja	en→ko	en→zh
M2M	DT	36.4	**28.5**	31.9	39.6	**28.5**	35.6	81.0	81.8	83.2
	DP	37.7	23.2	**33.2**	40.3	27.0	**36.8**	90.0	89.8	90.2
MBart	DT	**40.1**	20.8	27.1	41.1	20.3	28.6	**92.9**	88.8	87.3
	DP	38.3	20.2	26.8	**41.9**	21.3	30.5	91.2	**91.2**	**90.8**

Table 5: BLEU scores obtained with direct translation (DT) and detag-and-project (DP) using M2M and mBART models when evaluating markup split text (*XML BLEU*) and text only (*BLEU*). We also give the markup matching accuracies (*Markup Matching %*).

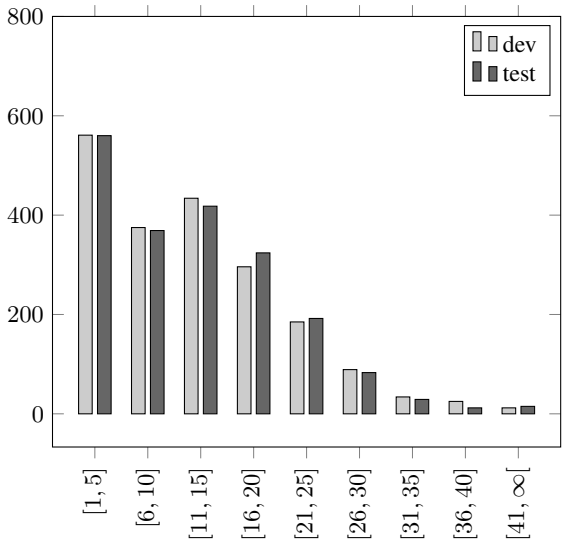

Figure 5: Length distributions of source segments.

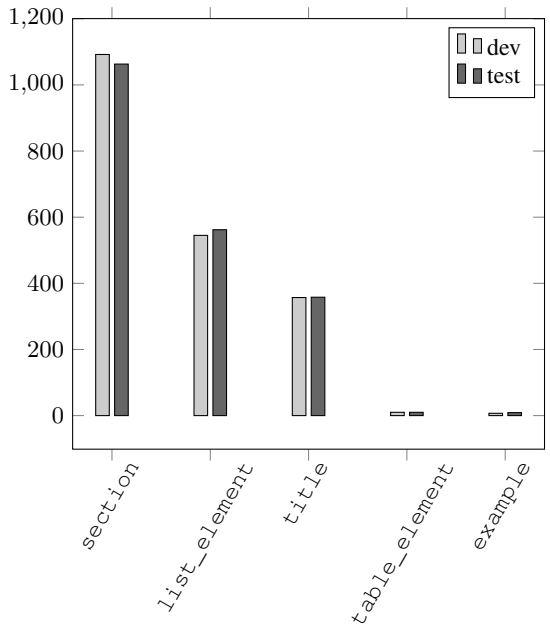

Figure 6: Distribution of textual element annotations.

A.3 Further Data Characteristics

Figure 5 shows the length distribution of English source segments. As typical for technical text, there is a high number of short sentences. In Figure 6 the distribution of textual element annotations is presented. It reflects, to some extent, the length distribution. The large proportion of list elements and titles accounts for shorter segments.

B Experiments and Results

B.1 Additional Evaluation

In addition to BLEU and chrF scores using SacreBLEU presented in Table 3 and 4, respectively, we present in Table 5, the scores obtained by using the evaluation metrics employed by (Hashimoto et al., 2019). Different from us, they report *XML BLEU* and *BLEU*. *XML BLEU* splits a translation containing inline markup into multiple parts relying on tags as split points. Note that the splitting takes place only if the markup structure in the translation and the reference match. In case of markup structure mismatch, the translation is treated as empty thereby penalizing the *XML BLEU* score. On the other hand *BLEU* is calculated by removing markup in the gold and translation, which is similar to our proposed *lex metrics* presented in Section 4.2. However, Hashimoto et al. (2019) use different tokenization methods compared to the ones implemented in sacreBLEU thus leading to different BLEU scores. Table 5 also contains the markup matching accuracy (*Markup Matching %*) which measures the number of examples with matching tags between the MT output and the reference translation.

Comparing *lex* scores in Table 3 and *BLEU* scores in Table 5, although the scores themselves are different and not directly comparable, the trends are similar where MBart is better than M2M only for English to Japanese translation and results for English to Korean translation are relatively lower compared to the two other translation directions. *XML BLEU* scores are usually lower than

BLEU scores because it penalizes translations with markup structure mismatch.

Markup Matching % for detag-project (DP) is typically higher than for direct translation (DT) because DP injects markup after translation whereas DT deals with markup during translation. Upon further investigation, we found that DT sometimes over- or under-generates markup spuriously leading to poorer markup matching accuracies. DP does not suffer from this issue. However, DP has another limitation where, if it is unable to align content with markup between the source and translation, markup injection does not take place. Therefore, DT will always result in translations containing markup unlike DP, even if the former may not inject tags with correct structure. This is the reason why *tag* scores in Tables 3 and 4 for DT models are higher than for DP models despite lower markup matching accuracies for the former.