Evaluation Dataset for Lexical Translation Consistency in Chinese-to-English Document-level Translation

Xiangyu Lei[1], **Junhui Li**[1*], **Shimin Tao**[2], **Hao Yang**[2]
[1]School of Computer Science and Technology, Soochow University, Suzhou, China
[2]Huawei Translation Services Center, Beijing, China
20215227062@stu.suda.edu.cn; lijunhui@suda.edu.cn
{taoshimin, yanghao30}@huawei.com

Abstract

Lexical translation consistency is one of the most common discourse phenomena in Chinese-to-English document-level translation. To better evaluate the performance of lexical translation consistency, previous researches assumes that all repeated source words should be translated consistently. However, constraining translations of repeated source words to be consistent will hurt word diversity and human translators tend to use different words in translation. Therefore, in this paper we construct a test set of 310 bilingual news articles to properly evaluate lexical translation consistency. We manually differentiate those repeated source words whose translations are consistent into two types: true consistency and false consistency. Then based on the constructed test set, we evaluate the performance of lexical translation consistency for several typical NMT systems.

Keywords: document-level neural machine translation, lexical translation consistency

1. Introduction

Sentence-level neural machine translation (NMT) has made great progress and development in recent years (Bahdanau et al., 2015; Wu et al., 2016; Vaswani et al., 2017). At the same time, researchers are paying more and more attention to document-level NMT (Maruf et al., 2022) which utilises inter-sentential context information in the document. Unlike sentence-level NMT, document-level NMT not only needs to pay attention to the dependencies between intra-sentences and inter-sentence, but also needs to consider much unique inter-sentence discourse phenomena, such as coreference, semantic coherence, and lexical cohesion. However, most previous studies on document-level NMT rarely try to model discourse phenomena explicitly, but incorporate discourse implicitly by using sentences in the wider-document context via different techniques in modeling context (Maruf et al., 2022).

Figure 1 shows an example of document-level Chinese-to-English translation. Comparing the output of sentence-level NMT with reference translation, we observe that source word 抗洪屋/resist-flood-house is consistently translated into *flood - resistant house* in reference while it appears as *flood fighting house*, *flood control house* and *flood resistance homes* in sentence-level NMT, respectively.[1] Meanwhile, context-aware document-level NMT (Bao et al., 2021; Li et al., 2023; Lupo et al., 2022), also cannot effectively solve such problem of lexical translation consistency, as it translates 抗洪屋/resist-flood-house into *flood prevention house*, *flood shelter*, and *flood resistant homes*. Although the meanings of these phrases are close, inconsistent translation may break the coherence of the text and tends to confuse the readers.

Source
<S1> . . . 最近 推出 了 一 款 “ 抗洪屋 ” . . .
<S10> “ 抗洪屋 ” 还 具有 特殊 的 通风口 . . .
<S12> . . . 但 他们 的 抗洪屋 可以 浮到 地面 五 米 以上 . . .

Reference
<S1> . . . recently unveiled a “ flood - resistant house ” . . .
<S10> the “ flood - resistant house ” also features special vents . . .
<S12> . . . but because their flood - resistant house can float up to five meters above ground . . .

Sentence-Level NMT
<S1> . . . recently launched a " flood fighting house " . . .
<S10> the " flood control house " also has a special ventilation outlet . . .
<S12> . . . however , their flood resistance homes can float over five meters above the ground . . .

Document-Level NMT
<S1> . . . recently introduced a " flood prevention house " . . .
<S10> the " flood shelter " also has a special flood valve . . .
<S12> . . . but their flood resistant homes can float up to five meters above ground level . . .

Figure 1: An example of document-level Chinese-to-English translation.

Following the idea of “one translation per discourse” (Merkel, 1996; Carpuat, 2009), many efforts have been devoted to increase the lexical translation consistency in document-level machine translation. In statistical machine translation, for example, relevant studies (Xiao et al., 2011; Garcia et al., 2014, 2017) propose differ-

*Corresponding author: Junhui Li.

[1] Following Lyu et al. (2021), we say repeated words w are consistently translated if their translations are same (stemmed if necessary).

LREC-COLING 2024, pages 6575–6581

ent post-editing approaches to re-translate those source words which appears multiple times in a document and are translated differently. Moving to NMT, studies (Kang et al., 2021; Lyu et al., 2021, 2022) explicitly model repeated source words in a source-side document and use different strategies to constrain their translations to be consistent. Alternatively, Zhang et al. (2023) improve translation consistency via document-level translation repair.

However, constraining translations of all repeated source words to be consistent will hurt word diversity, especially for newswires. As shown in the source-target document pair $(\mathcal{X}, \mathcal{Y})$ in Figure 2, the translations of repeated source words can be further categorized into three groups:

- **True consistency.** The translations of a repeated source word in $\mathcal{Y}$ are consistent, and it is *not* acceptable if they are inconsistent. For example, the translations of source word 供应/supply are of this type.

- **False consistency.** Although the translations of repeated source word in $\mathcal{Y}$ are consistent, it would not change the meaning and hurt cohesion if they are inconsistent. For example, although repeated source word 提高/improve is consistently translated into *improve* in $\mathcal{Y}$, it is totally acceptable if it is translated into *improve* once and *increase* another once.

- **Inconsistency.** The translations of a repeated word in $\mathcal{Y}$ are inconsistent. For example, repeated source word 增强/increase is translated into *increase* and *enhance*.

Therefore, when evaluating a NMT system on the performance of lexical translation consistency, no matter sentence-level or document-level, it is critical to focus on those repeated words whose translations are truly consistent (e.g., 供应/supply in Figure 2), rather than those repeated words whose translations are of either false consistency or inconsistency (e.g., 提高/improve and 增强/increase in Figure 2).

To this end, this paper proposes a test set to evaluate lexical translation consistency in Chinese-to-English document-level translation. The test set contains 310 parallel documents, each of which consists of multiple parallel sentences.[2] In each source-side document, we provide several pairs of repeated source words whose translations are truly consistent. Based on the constructed test set, we then evaluate the performance of lexical translation consistency for several NMT systems.

Source $\mathcal{X}$

<S6> 中国 将 努力 提高 能源 供应 保障 能力 ， 增强 能源 领域 供应链 的 安全性 和 稳定性 。

<S8> 为 提高 能源 生产 和 供应 保障 能力 ， 将 积极 推进 智能化 煤矿 核准 建设 ， 推动 在 建 煤矿 尽快 投产 达产 ， 全面 增强 煤炭 安全 增产 保供 的 能力 和 韧性 。

Target y

<S6> china will make efforts to improve the ability to ensure energy supply , increase the security and stability of the supply chain ...

<S8> to improve the ability to ensure energy production and supply , ... , and enhance the ability to safely increase coal production ...

Figure 2: Example of repeated words belonging to true consistency (e.g., 供应/supply), false consistency (e.g., 提高/improve), and inconsistency (e.g., 增强/increase).

2. Test Set Construction

2.1. Background

By studying the distribution of discourse phenomena in three different genres, including news, talks, and subtitles, Kang et al. (2021) reveal that lexical inconsistency is the most serious issues in document-level Chinese-to-English translation. For example, in the genre of news, lexical consistency accounts for 43.9% of all errors while tense consistency and pronoun translations account for 24.5% and 9.2%, respectively.

It is widely acknowledged that automatic evaluation metrics, such as BLEU, lack sensitivity in assessing lexical translation consistency. Previous studies by Bawden et al. (2018) and Voita et al. (2019) have introduced contrastive test sets focusing on coherence and cohesion. However, their methodology evaluates model performance by comparing generation probabilities of positive and negative examples. In real-world scenarios, the diversity of discourse phenomena poses a challenge, making it impractical to encompass all potential variations.

In this paper we focus on the genre of news and choose the bilingual news from China Daily[3] as the corpus. These news articles, originally composed in English from diverse media outlets such as The Guardian, Xinhua, Mental Floss, etc., have been professionally translated into Chinese. The chosen topics span politics, business, entertainment, lifestyle, health, and more. During the collection of bilingual news, documents lacking sentence-level alignment are excluded. Additionally, repeated words on the source side are manually annotated. The evaluation of the model's ability to maintain lexical translation consistency involves assessing whether the model consistently translates these

[2] We will release the test set soon.

[3] http://language.chinadaily.com.cn/

repeated words.

2.2. Construction

The construction of our test set mainly consists of three steps: word segmentation, word alignment, and manual annotation.

Word segmentation. We segment the source-side Chinese sentences with Joint-Parser (Hou et al., 2021) and tokenize the target-side English sentences with SuPar (Zhang et al., 2020). While infrequent, inconsistencies in Chinese word segmentation results are possible. For example, sequence 美联储/America-union-deposit appears multiple times in a document while it is segmented into 美/America 联储/union-deposit and 美 联 储/America-union-deposit in different sentences. In such cases, we address these inconsistencies through manual correction.

Word alignment. We use awesome-align (Dou and Neubig, 2021) to obtain word alignment between the bilingual sentences. Our focus lies in annotating the translation consistency of source-side words. Specifically, we concentrate on one-to-one and one-to-many alignments, indicating that a single source word aligns with either one or multiple target words.

Manual annotation. We train two linguistics students bilingual in Chinese and English. Given a document pair $(\mathcal{X}, \mathcal{Y})$ and the word alignment result $\mathcal{A}$, we extract N lexical chains $\mathcal{C} = \{C^i\}|_{i=1}^{N}$. Each lexical chain $C^i = \{w^i, t^i\left(a_l^i, b_l^i\right)|_{l=1}^{L}\}$ records all positions of word w^i that appears L times ($L \geq 2$) in $\mathcal{X}$ and their translations are same, where t^i is their translation which may consist of a single, or multiple words, a and b indicate the sentence index and word index of a position, respectively.[4] Please note that our lexical chains are different from those defined in Lyu et al. (2022) and Zhang et al. (2023), where they do not require the translations in $\mathcal{Y}$ are same. For each lexical chain $\mathcal{C}^i$ with L items, an annotator is asked if any of them can be translated differently without changing the meaning and hurting cohesion. For example, if the j-th item in the chain can be differently translated from the others, then we say it is acceptable if the pair of source word $\left(a_l^i, b_l^i\right), \left(a_j^i, b_j^i\right)$ is translated inconsistently. Otherwise, they should be translated consistently.

2.3. Statistics

The test set contains a total of 310 bilingual news articles from China Daily, with a total of 5,051 sentences, and an average of 16 sentences per article. Each of the two students annotate 176 bilingual news articles, with 42 articles overlapping for the purpose of calculating inter-annotator agreement (IAA). The resulting IAA F1 score is 0.96, suggesting a high level of agreement. Further details on the IAA calculation can be found in Appendix A.

True Consistency	False Consistency	All
17,292 (56.4%)	13,375 (43.6%)	30,667

Table 1: Statistics on the repeated source words which are of true and false consistency.

NOUN	VERB	ADJ/ADV	OTHERS
14,349 (83.0%)	327 (1.9%)	111 (0.6%)	2,505 (14.5%)

Table 2: Statistics over different POS tags.

Given the varying frequencies of repetitions for source-side words, we proceed to calculate pairs of repeated source words. Specifically, if a source word w appears k times in the source-side document $\mathcal{X}$, we say there are $\frac{k\times(k-1)}{2}$ pairs. Among all the pairs of repeated source words whose translations are consistent in target-side document, Table 1 shows that almost 44% of them are of false consistency. This suggests that these words can be translated differently without changing the meaning and hurting cohesion.

Among the 17,292 pairs of repeated source words which are of true consistency, Table 2 shows the translation consistency of repeated source words with different POS (part-of-speech) tags. From it we observe that repeated nouns (with POS tags of NN, NR, and NT) are more like to be translated consistently than verbs (with POS tags of VV, VE, and VC), adjectives/adverbs (with POS tag of VA/AD), and others. This is consistent with the findings in Kang et al. (2021) and Guillou (2013).

For distance perspective, we define the inter-sentence distance of each pair as d. Table 3 shows the statistics over different distances. It shows that the size of repeated source word pairs decreases when the distance increases from 1. Moreover, repeated source word pairs with distance of ≥ 5 account for 50%, indicating that consistent translation can be frequently observed with long context.

3. Experimentation

In order to verify the usefulness of our proposed test set, we compare the performance of several typical NMT systems.

3.1. Experimental Settings

Considering that the current popular document-level NMT systems may use local or global context for modeling, we test lexical translation consistency on the constructed test set with the following four NMT models.

[4] We only consider content words.

$d = 0$	$d = 1$	$d = 2$	$d = 3$	$d = 4$	$d >= 5$
945 (5.5%)	2,571 (14.9%)	2,013 (11.6%)	1,729 (10.0%)	1,424 (8.2%)	8,610 (49.8%)

Table 3: Statistics over different distances.

Model	**NOUN**		**VERB**		**ADJ/ADV**		**OTHERS**	
	L-All	L-Anno	L-All	L-Anno	L-All	L-Anno	L-All	L-Anno
Sent-Transformer	66.34	69.10	52.89	56.57	65.61	**58.56**	68.34	66.95
Doc-Transformer	65.83	69.51	50.74	60.55	62.85	42.34	63.52	65.15
G-Transformer	66.81	70.51	53.53	**62.39**	63.53	49.55	67.95	70.94
CVAE-Transformer	**67.62**	**72.69**	**55.36**	58.72	**65.62**	57.66	**71.14**	**71.66**

Table 4: Performance (LTCR) on the constructed test set from POS perspective.

Model	**0**		**1**		**2**		**3**		**4**		**>=5**	
	L-All	L-Anno	L-All	L-Anno	L-All	L-Anno	L-All	L-Anno	L-All	L-Anno	L-All	L-Anno
Sent-Transformer	73.98	68.68	64.48	67.37	64.82	67.76	64.52	68.83	65.91	71.42	65.33	68.41
Doc-Transformer	72.34	65.61	63.48	66.04	63.62	67.11	63.79	67.96	64.33	68.82	64.25	70.00
G-Transformer	73.05	69.95	66.19	70.36	63.68	68.95	64.45	70.27	64.76	70.44	66.11	70.58
CVAE-Transformer	**74.50**	**72.06**	**66.72**	**71.96**	**66.49**	**70.54**	**67.08**	**70.97**	**67.69**	**73.24**	**66.54**	**72.71**

Table 5: Performance (LTCR) on the constructed test set from distance perspective.

- **Sent-Transformer** (Vaswani et al., 2017), which is a context-agnostic model.
- **Doc-Transformer** (Zhang et al., 2018), which models two previous source-side sentences when translating current sentences.
- **G-Transformer** (Bao et al., 2021), which view document-level translation as a sequence-to-sequence generation task. It translate each sentence by modeling global source-side document and previous sentences in the target-side.
- **CVAE-Transformer** (Lyu et al., 2022), which models global source-side document when translating current sentences. Specifically, it aims to improve lexical translation consistency by modeling source-side lexical chains of repeated source words.

Training data and strategy. The training set, sourced from LDC, consists of 2 million sentence pairs. The document-level parallel corpus is a subset of the full training set, consisting of 66K documents with a total of 0.8 million sentence pairs. For pre-training sentence-level parameters, we employ sentence parallel corpora, while document parallel corpora are utilized to fine-tune document-level parameters. Additional information can be found in Appendix B.

Evaluation metrics. We report LTCR (Lexical Translation Consistency Ratio) score (Lyu et al., 2021) for evaluating lexical translation consistency. Specifically, we report LTCR scores based on all source-side repeated words (L-All for short) and the annotated repeated words that are of true consistency (L-Anno for short), respectively. Moreover, we report case-insensitive BLEU calculated by the SacreBLEU (Post, 2018).

Model	**BLEU**	**L-All**	**L-Anno**
Sent-Transformer	17.77	65.63	68.48
Doc-Transformer	18.12	64.51	68.53
G-Transformer	18.04	66.00	70.28
CVAE-Transformer	**18.33**	**67.20**	**72.18**

Table 6: Performance (BLEU and LTCR) on the constructed test set.

3.2. Results

Main results. Table 6 shows the performance on the test set. First, by modeling global document-level context, both G-Transformer and CVAE-Transformer obtain improvement in both BLEU and LTCR scores while CVAE-Transformer achieves the best performance. However, by modeling local document-level context, Doc-Transformer does not achieves better lexical consistency than the sentence-level baseline. Second, the performance trend of BLEU and LTCR scores is not always consistent. For example, Doc-Transformer achieves similar performance in BLEU (i.e., 18.12 vs. 18.04) than G-Transformer, but the latter gets better performance in LTCR scores. Third, the performance trend for the two LTCR scores L-All and L-Anno is not always consistent neither. For example, Doc-Transformer obtains lower L-All but higher L-Anno score than the baseline.

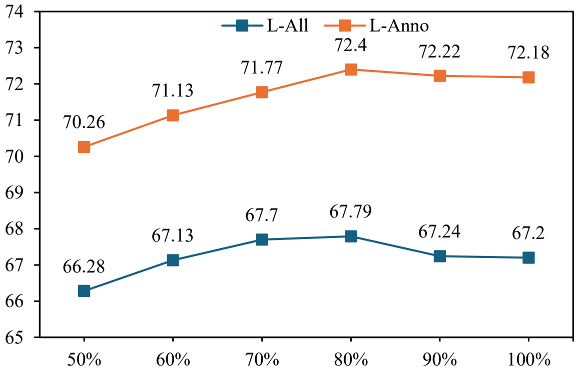

Figure 3: Performance trend in L-All and L-Anno scores for CVAE-Transformer across different sample sizes.

Performance from POS perspective. Table 4 details the performance from the POS perspective. First, we observe that repeated nouns are more like to be translated consistently than verbs, adjectives, and adverbs, whose translations are more flexible. Second, except repeated words of ADJ/ADV, all four NMT models achieves higher L-Anno scores than L-All scores. This indicates that repeated words of true consistency are indeed translated more consistently than repeated words of false consistency and inconsistency.

Performance from distance perspective. Table 5 shows the performance from the distance perspective. First, for all NMT models we observe that L-All score is higher than L-Anno score when the distance is 0. This indicates that NMT models tend to translate repeated words within a sentence consistently. For repeated words appearing in different sentences (i.e., $d > 0$), for all NMT models the L-Anno score is higher than L-All. Second, we also observe that for all NMT models the L-All/L-Anno scores are similar for different distances.

Adequacy discussion. To better illustrate the adequacy, we include a performance trend analysis that takes into account different sizes of the test sets. Figure 3 shows the performance trend in L-All score and L-Anno score for CVAE-Transformer across different sample sizes. By comparing the trends in the two scores, it can be observed that as the sample size changes, the L-All and L-Anno scores tend to stabilize. However, the variation in the L-Anno score is more stable compared to L-All score, and L-Anno score provides a more intuitive reflection of lexical translation consistency.

4. Conclusion

In this paper we have presented a test set for properly evaluating lexical translation consistency in Chinese-to-English document-level translation. Based on the test set, we evaluate the performance of lexical translation consistency for four NMT systems. Our analysis show that the constructed test set could effectively evaluate the lexical translation consistency in document-level translation. In future work, we would like to evaluate the lexical translation consistency on our test set with large language model (LLM)-based document-level NMT models (Li et al., 2024; Lyu et al., 2024).

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive feedback. This work was supported by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR*.

Guangsheng Bao, Yue Zhang, Zhiyang Teng, Boxing Chen, and Weihua Luo. 2021. G-transformer for document-level machine translation. In *Proceedings of ACL*, pages 3442–3455.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceeding of NAACL*, pages 1304–1313.

Marine Carpuat. 2009. One translation per discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 19–27.

Ziyi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceeding of EACL*, pages 2112–2118.

Eva Martínez Garcia, Carles Creus, Cristina Espana-Bonet, and Lluís Màrquez. 2017. Using word embeddings to enforce document-level lexical consistency in machine translation. *Prague Bulletin of Mathematical Linguistics*, 108:85–96.

Eva Martínez Garcia, Cristina Espana-Bonet, and Lluís Màrquez. 2014. Document-level machine translation as a re-translation process. *Procesamiento del Lenguaje Natural*, 53:103–110.

Liane Guillou. 2013. Analysing lexical consistency in translation. In *Proceedings of DiscoMT*, pages 10–18.

Yang Hou, Houquan Zhou, Zhenghua Li, Yu Zhang, Min Zhang, Zhefeng Wang, Baoxing Huai, and Nicholas Jing Yuan. 2021. A coarse-to-fine labeling framework for joint word segmentation, pos tagging, and constituent parsing. In *Proceeding of CoNLL*, pages 290–299.

Xiaomian Kang, Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2021. Enhancing lexical translation consistency for document-level neural machine translation. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 21(3):1–21.

Yachao Li, Junhui Li, Jing Jiang, Shimin Tao, Hao Yang, and Min Zhang. 2023. P-Transformer: Towards Better Document-to-Document Neural Machine Translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3859–3870.

Yachao Li, Junhui Li, Jing Jiang, and Min Zhang. 2024. Enhancing document-level translation of large language model via translation mixed-instructions. *CoRR*, abs/2401.08088.

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. 2022. Divide and rule: Effective pre-training for context-aware multi-encoder translation models. In *Proceedings of ACL*, pages 4557–4572.

Xinglin Lyu, Junhui Li, Zhengxian Gong, and Min Zhang. 2021. Encouraging lexical translation consistency for document-level neural machine translation. In *Proceeding of EMNLP*, pages 3265–3277.

Xinglin Lyu, Junhui Li, Shimin Tao, Hao Yang, Ying Qin, and Min Zhang. 2022. Modeling consistency preference via lexical chains for document-level neural machine translation. In *Proceeding of EMNLP*, pages 6312–6326.

Xinglin Lyu, Junhui Li, Yanqing Zhao, Min Zhang, Daimeng Wei, Shimin Tao, Hao Yang, and Min Zhang. 2024. Dempt: Decoding-enhanced multi-phase prompt tuning for making llms be better context-aware translators. *CoRR*, abs/2402.15200.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2022. A survey on document-level neural machine translation: Methods and evaluation. *ACM Computing Surveys*, 54:45:1–45:36.

Magnus Merkel. 1996. Consistency and variation in technical translation: a study of translators' attitudes. In *Proceedings of Unity in Diversity, Translation Studies Conference*, pages 137–149.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of WMT*, pages 186–191.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceeding of NIPS*, pages 5998–6008.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of ACL*, pages 1198–1212.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Tong Xiao, Jingbo Zhu, Shujie Yao, and Hao Zhang. 2011. Document-level consistency verification in machine translation. *Machine Translation Summit XIII*, 13:131–138.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, FeiFei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of EMNLP*, page 533–542.

Yu Zhang, Houquan Zhou, and Zhenghua Li. 2020. Fast and accurate neural crf constituency parsing. In *Proceeding of IJCAI*, pages 4046–4053.

Zhen Zhang, Junhui Li, Shimin Tao, and Hao Yang. 2023. Lexical translation inconsistency-aware document-level translation repair. In *Findings of ACL*, page 12492–12505.

A. IAA F1 Score

We calculate IAA (Inter-Annotator Agreement) F1 score as:

$$\text{Percision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{IAA F1 score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

where TP is the number of samples correctly annotated as consistent, FP is the number of samples incorrectly annotated as consistent, and FN is the number of samples incorrectly annotated as inconsistent.

B. Experiment Dateset and Training Strategy

The sentence-level training set consists of LDC2002E18, LDC2003E07, LDC2003E14, news part of LDC2004T08 and the document-level training set from LDC2002T01, LDC2004T07, LDC2005T06, LDC2005T10, LDC2009T02, LDC2009T15, LDC2010T03. The development set consists of NIST 2006. The test set consists of NIST 2002, 2003, 2004, 2005 and 2008.

In the pre-training stage, we train the sentence-level models on sentence-level training set with 2M sentence pairs. And in the fine-tuning stage, we train the document-level models on using document-level training set with 66K documents.

For all translation models, the hidden size and the filter size are set to 512 and 2,048, respectively. The number of heads in multi-head attention is set to 8. The numbers of layers in the encoder and the decoder are set to 6. And we use Adam with β_1=0.9 and β_2=0.98 for optimization.