

Improving Lexically Constrained Neural Machine Translation with Source-Conditioned Masked Span Prediction

Gyubok Lee* Seongjun Yang* Edward Choi

Graduate School of AI, KAIST

{gyubok.lee, seongjunyang, edwardchoi}@kaist.ac.kr

Abstract

Accurate terminology translation is crucial for ensuring the practicality and reliability of neural machine translation (NMT) systems. To address this, lexically constrained NMT explores various methods to ensure pre-specified words and phrases appear in the translation output. However, in many cases, those methods are studied on general domain corpora, where the terms are mostly uni- and bi-grams (>98%). In this paper, we instead tackle a more challenging setup consisting of domain-specific corpora with much longer n-gram and highly specialized terms. Inspired by the recent success of masked span prediction models, we propose a simple and effective training strategy that achieves consistent improvements on both terminology and sentence-level translation for three domain-specific corpora in two language pairs.

1 Introduction

Despite its recent success in neural machine translation (NMT) (Wu et al., 2016; Johnson et al., 2017; Barrault et al., 2020), delivering correct terms in the translation output is still a vital component for high-quality translation. This concern becomes more salient in domain-specific scenarios, such as in legal documents, where generating correct and consistent terminology is key to ensuring the practicality and reliability of machine translation (MT) systems (Chu and Wang, 2018; Exel et al., 2020).

To address this, lexically constrained NMT works have proposed various methods to preserve terminology in translations as lexical constraints with or without the help of a term dictionary at test time. In most lexically constrained NMT setups, datasets and terms used for training and evaluating the methods are extracted from WMT news corpora (Dinu et al., 2019; Susanto et al., 2020; Chen

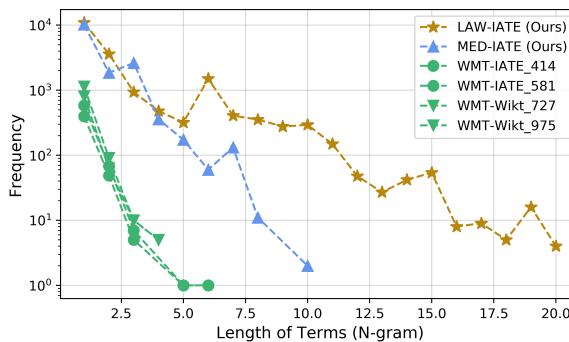


Figure 1: The frequency of terms sorted by n-gram between Dinu et al. (2019)’s and our test splits. While the terms in WMT De-En are mostly uni- or bi-grams, our setup contains heavy-tailed n-gram distributions with more quantity and diversity in terminology.

et al., 2020). Since the terms, regardless of their source, can only be utilized as long as they exist in the corpus, the term coverage solely depends on the choice of the corpus. By analyzing the previous setups carefully, we discover that the terms found in WMT are mostly uni- or bi-grams (see Figure 1) and highly colloquial (see Table 1 for the top 10 most frequent terms). These leave the question of whether the previous methods are effective in domain-specific scenarios where accurate terminology translation is truly vital.

In this paper, inspired by the recent masked span prediction models, which have demonstrated improved representation learning capability of contiguous words (Song et al., 2019b; Joshi et al., 2019; Lewis et al., 2020; Raffel et al., 2020), we propose a simple yet effective training scheme to improve terminology translation in highly specialized domains. We specifically select two highly specialized domains (*i.e.*, law and medicine) which contain domain-specific terminologies to address more challenging and realistic setups, in addition to applying it to both typologically similar and dissimilar pairs of languages (German-English (De→En) and Korean-English (Ko→En)). Thanks to its sim-

* equal contributions

plicity, the proposed method is compatible with any autoregressive Transformer-based model, including ones capable of utilizing term dictionaries at training or test time. In domain-specific setups where longer n-gram terms are pervasive, our method demonstrates improved performance over the standard maximum likelihood estimation (MLE) approach in terms of terminology and sentence-level translation quality. Our code and datasets are available at https://github.com/wns823/NMT_SSP.

2 Background

Lexically constrained NMT We could group lexically constrained NMT methods into two streams: *hard* and *soft*. The hard approaches aim to force all terminology constraints to appear in the generated output. The methods include replacing constraints (Crego et al., 2016), constrained decoding (Hokamp and Liu, 2017; Chatterjee et al., 2017; Post and Vilar, 2018; Hasler et al., 2018), and additional attention heads for external supervision (Song et al., 2020). Although those approaches are reliable and widely used in practice, they typically require a pre-specified term dictionary and an extra candidate selection module if there are multiple matching candidates for a single term (see caption in Table 2).

Several soft methods address this problem without the help of a term dictionary, one of which is training on both constraint pseudo-labeled (with statistical MT) and unlabeled data (Song et al., 2019a). More recently, Susanto et al. (2020) and Chen et al. (2020) proposed methods that do not assume any word alignment or dictionary supervision at training time to handle unseen terms at test time. For their flexibility, we choose them as our baselines. As discussed in Section 1, most previous methods are trained and evaluated on general domain corpora. In this work, we instead tackle highly specialized domain-specific corpora such as law and medicine, where the terms are much longer and often rare.

Domain-specific NMT Another line of research related to our problem is domain-specific NMT, where difficulties arise from both a limited amount of parallel data and specialized lexicons. Similar to the hard approaches in lexically constrained NMT, several works rely on domain-specific dictionaries (Zhang and Zong, 2016a; Hu et al., 2019; Thompson et al., 2019; Peng et al., 2020) when generating translations, but they are also prone to the same is-

Dinu et al. (2019)'s dataset	
late-414	gold(15), CDU(13), bridge(12), China(11), Syria(11), night(11), campaign(11), generation(9), month(7), Iraq(7)
late-581	gold(26), doping(23), CDU(19), sport(17), US(15), bridge(14), Syria(13), campaign(13), China(11), night(11)
Wikt-727	percent(61), police(50), Thursday(41), Putin(19), five(17), September(14), Venus(13), later(12), Tuesday(11), less(11)
Wikt-975	percent(61), police(59), Thursday(44), Putin(24), old(21), September(21), five(16), swimmer(14), later(13), Venus(13)
Our dataset	
Law (De-En)	Council(706), Regulation(521), Commission(481), Union(478), Treaty(345), Official Journal(319), Member State(283), proposal(239), on a proposal from the Commission(229), market(181)
Medical (De-En)	injection(469), water(275), water for injection(270), patient(269), infusion(258), solution for infusion(226), sodium(159), distribution(127), volume of distribution(125), treatment(120)
Law (Ko-En)	si(451), official(445), public official(436), member(436), term of office(367), gu(265), education(209), period(180), term(180), management(156)

Table 1: Top 10 most frequent terms in Dinu et al. (2019)'s and our test splits. Numbers in parenthesis indicate the frequency of terms in each data. As shown in the two tables, all top 10 terms in the WMT corpus are unigrams, while there are longer terms (up to 6-grams) in the domain-specific corpora. Furthermore, compared to WMT, the terms in the domain-specific corpora are more specialized for their corresponding domains.

sues. Other domain-specific NMT methods include unsupervised lexicon adaptation (Hu et al., 2019), synthetic parallel data generation with monolingual data (Sennrich et al., 2016a), and multi-task learning that combines language modeling and translation objectives (Gulcehre et al., 2015; Zhang and Zong, 2016b; Domhan and Hieber, 2017). Our method is a form of multi-task learning by utilizing both the source and target language text for an additional task, while the previous works mostly use only the target language text.

Span-based Masking Span-based masking is to predict the spans of masked tokens, as opposed to individual token predictions in BERT (Devlin et al., 2019). With this training objective, the model showed improved performance on span-level tasks including question answering and coreference resolution (Joshi et al., 2019). Concurrently, autoregressive sequence-to-sequence pre-trained models also utilized span-based masking as their objectives and demonstrated its effectiveness in many downstream tasks (Song et al., 2019b; Lewis et al., 2020; Raffel et al., 2020). Similar to theirs, our training scheme takes advantage of autoregressive

span-based prediction but we condition on both the source language and the previous non-masked target language tokens.

3 Approach

3.1 Source-Conditioned Masked Span Prediction

We posit that adopting auxiliary span-level supervision in generation can benefit both short and long terminology and sentence-level translation. We, therefore, propose an extra span-level prediction task in translation—namely, source-conditioned masked span prediction (SSP). Different from the recent sequence-to-sequence pre-trained models (Song et al., 2019b; Lewis et al., 2020; Raffel et al., 2020), our approach applies span masking only on the target side. By conditioning on the full context of the source language and the previous non-masked target language tokens (due to autoregressive decoding), the model is forced to predict the spans of missing tokens given fully referenced information in the encoder and partially in the decoder.

Span masking We follow the masking procedure proposed in SpanBERT (Joshi et al., 2019), where we first sample the length of spans from a clamped geometric distribution ($p=0.2$, $\max=10$) and then corrupt 80% of masked tokens with [MASK], 10% with random tokens, and 10% unchanged. We set the corruption ratio to 50%.

Multi-task Learning As our training scheme consists of two objectives (*i.e.* translation and masked span prediction), we define the total training objective as follows. Let θ be the model parameter and \mathbb{C} be the term-matched corpus where each sentence contains at least one or more terms. The first objective, translation, is to maximize the likelihood of the conditional probability of y :

$$p_{\theta}(y|x) = \prod_{t=1}^{T+1} p_{\theta}(y_t|y_{0:t-1}, x), \quad (1)$$

where $y = (y_1, \dots, y_T)$ is the target ground-truth (GT) sentence with length T and $x = (x_1, \dots, x_S)$ is the source sentence with length S . For the SSP objective, we first corrupt random spans of y until the corruption ratio, resulting in \tilde{y} . Then we autoregressively predict the masked tokens \tilde{y} while

		# Sent.	Avg. words per sent.	# Terms	Avg. terms per sent.	# Unique terms
Law (De→En)	SRC	447,410	27.46	1,677,852	2.33	25,460
	TRG	(441K/3K/3K)	30.77			27,755
Medical (De→En)	SRC	494,316	19.01	1,494,269	1.34	8,633
	TRG	(488K/3K/3K)	20.25			8,990
Law (Ko→En)	SRC	93,240	16.52	353,894	3.52	2,354
	TRG	(87K/3K/3K)	34.56			2,733

Table 2: Statistics of the filtered corpus and matched terms. Note that *# unique terms* in the source (SRC) and target (TRG) languages are not the same. For instance, “Arzneimittel” can translate into multiple forms—“pharmaceutical products”, “drug”, “medicinal product”, etc.—depending on the context.

conditioned on both \tilde{y} and x :

$$p_{\theta}(\tilde{y}|y, x) = \prod_{t=1}^{T+1} m_t p_{\theta}(y_t|\tilde{y}_{0:t-1}, x), \quad (2)$$

where $m_t = 1$ means y_t is masked.

Finally, we simultaneously optimize the joint loss:

$$\mathcal{L}_{total} = -\frac{1}{|\mathbb{C}|} \sum_{\substack{(x,y) \sim \mathbb{C}, \\ \tilde{y} \sim C(y)}} \log p_{\theta}(y|x) + \gamma \log p_{\theta}(\tilde{y}|y, x), \quad (3)$$

where C is a span-level corrupter and γ is a task coefficient that weights the relative contribution of SSP.

4 Experiments

4.1 Setup

Data We use De-En legal and medical domain corpora from OPUS¹ (Tiedemann, 2012) and the De-En bilingual term dictionary from IATE². Terms in different languages are aligned via term IDs. For the typologically distant pair of languages, we use the Ko-En legal domain corpus available on AI Hub³, and the manually processed bilingual term dictionary downloaded from the Korea Legislation Research Institute (KLRI) website⁴. In cases where

¹<http://opus.nlpl.eu/>

²<https://iate.europa.eu>

³<https://www.aihub.or.kr/aidata/87>

⁴https://www.moleg.go.kr/board.es?mid=a1050400000&bid=0010&act=view&list_no=43927&nPage=2

Model	Law (De→En)						Law (Ko→En)							
	BLEU	Term% (↑)				LSM-3 (↑)		BLEU	Term% (↑)				LSM-3 (↑)	
		1-gram	2-gram	2>micro	2>macro	2>micro	2>macro		1-gram	2-gram	2>micro	2>macro	2>micro	2>macro
GU19	70.64	94.36	92.33	73.31	45.25	86.22	74.92	51.31	81.47	76.51	58.15	38.51	69.41	62.52
VASWANI17	75.24	95.80	93.87	80.29	55.31	89.71	79.77	53.01	84.97	81.29	65.79	54.55	74.29	70.56
+SSP	75.44	96.01	94.08	81.52	58.79	90.81	82.50	53.80	85.84	83.94	66.84	58.15	75.71	69.82
CHEN20	74.19	95.55	94.08	80.63	54.73	89.80	80.89	53.08	85.49	83.25	65.51	52.49	74.53	67.81
+SSP	75.24	95.92	94.50	81.31	56.33	90.40	81.91	53.32	85.63	82.10	66.19	56.50	76.02	72.27

Table 3: (*Without dictionary*) Results on legal domain corpora (De→En and Ko→En) without terminology guidance at test time. VASWANI17 combined with our training objective (Eq.(3)) outperforms other methods in most cases. Note that GU19 is a non-autoregressive model, therefore not applicable to our proposed method. Higher Term% and LSM-3 mean better performance.

Model	Law (De→En)						Law (Ko→En)							
	BLEU	Term% (↑)				LSM-3 (↑)		BLEU	Term% (↑)				LSM-3 (↑)	
		1-gram	2-gram	2>micro	2>macro	2>micro	2>macro		1-gram	2-gram	2>micro	2>macro	2>micro	2>macro
SUSANTO20	62.20	94.38	92.95	82.06	64.06	94.93	92.14	50.56	81.67	76.74	58.47	38.66	69.63	62.60
CHEN20	73.05	96.64	93.29	78.73	51.47	90.00	80.29	52.60	84.74	83.94	67.33	59.53	75.59	74.54
+SSP	74.72	97.15	95.95	84.67	57.48	93.94	83.62	53.38	95.86	94.92	88.58	79.34	94.17	91.48

Table 4: (*With dictionary*) Results on legal domain corpora when the GT terms are provided at test time. +SSP consistently shows improvements over its MLE counterparts. Contrary to the previous findings (Susanto et al., 2020; Chen et al., 2020), the models do not show improved BLEU scores compared to those in Table 3. We argue that providing terms at test time is indeed helpful for terminology generation, but it can often hinder the generation of fluent text. This becomes more apparent in our non-autoregressive setup.

one term translates into multiple terms, we consider all possible pairs to maximize the number of sentence and term matches.

To avoid trivial matches between the parallel sentences and terms, we filtered out terms that are less than four characters long and longer than 20 grams. Sentences that do not contain any term are also removed. The statistics of the datasets are reported in Table 2. More details about the preprocessing steps are in Appendix A.1.

For data splitting, we developed a new data splitting algorithm that considers the same distribution of n-grams across each data split. We use 3,000 sentences for valid and test sets in case of high redundancy in certain corpora, while previous works that utilize OPUS use only 2,000 (Koehn and Knowles, 2017; Müller et al., 2020). It is important to note that all the sentences in our data splits are matched with domain-specific terms (i.e. at least one or more terms exist in each sentence) following the style of Dinu et al. (2019). The pseudo-code for the terminology-aware data split algorithm is in Appendix B.

Baselines We compare our method on two recent lexically constrained NMT models of different natures: autoregressive (Chen et al., 2020) and

non-autoregressive (Susanto et al., 2020), but both can operate with or without a term dictionary at test time. We refer to them as CHEN20 and SUSANTO20, respectively. +SSP indicates models trained with our proposed training scheme, while no indication is the standard MLE method. A base Transformer (Vaswani et al., 2017), denoted as VASWANI17, and a Levenshtein Transformer (Gu et al., 2019), denoted as GU19, are also reported to compare the relative performance between models. SUSANTO20 without a dictionary is equivalent to GU19.

Evaluation We use SacreBLEU⁵ (Post, 2018) for measuring translation quality. For terminology translation, we use *term usage rate* for both short (≤ 2 -grams) and long (> 2 -grams) terms. Term usage rate (Term%) is the number of generated terms divided by the total number of terms (Dinu et al., 2019; Susanto et al., 2020). Specifically for evaluating long terms, we report both the macro and micro averages due to the heavy-tailed nature of n-grams. In addition, although exact term translation is the primary objective for terminology translation, due to its harshness, evaluating models only with Term% may not fully describe the models’

⁵<https://github.com/mjpost/sacrebleu>

Model	Medical (De→En)						
	BLEU	Term% (↑)				LSM-3 (↑)	
		1-gram	2-gram	2>micro	2>macro	2>micro	2>macro
<i>Without dictionary</i>							
GU19	70.85	93.83	91.24	77.46	53.66	86.15	75.21
VASWANI17	76.31	94.22	90.80	79.82	61.03	87.11	80.48
+SSP	76.87	94.36	91.31	80.63	53.68	88.01	74.76
CHEN20	74.84	94.29	90.61	79.42	68.37	87.13	84.64
+SSP	76.72	94.61	90.42	80.41	68.03	88.08	83.04
<i>With dictionary</i>							
SUSANTO20	62.20	91.01	92.64	88.09	67.22	95.46	94.27
CHEN20	72.84	94.40	93.58	83.77	67.98	89.95	86.70
+SSP	75.50	95.86	94.92	88.58	79.34	94.17	91.48

Table 5: Results on the medical domain dataset (De→En).

behavior. Therefore, we also evaluate each model in terms of partial n-gram matches, which is explained in the next paragraph. All evaluations are conducted with a beam size of 5.

Partial N-gram Match Inspired by the longest common substring problem (Gusfield, 1997), we devised a partial n-gram match score for evaluating long terminology—*longest sub n-gram match* (LSM) score. Formally, let the generated target sentence be $\hat{y} = (\hat{y}_1, \dots, \hat{y}_T)$ and the matched terms for the target ground truth (GT) sentence y be $y' = \bigcup_{i=1}^N (y'_{i1}, \dots, y'_{il})$, where N is the number of GT terms in y and l is an arbitrary n-gram length for i -th term. Then, LSM is defined as the ratio of the longest n-gram overlap divided by l . As too many overlaps can occur at the uni and bi-gram levels, we only calculate LSM for long terminology, which means the least overlap has to be greater than or equal to 3 grams, all else being zero, therefore denoted as LSM-3.

4.2 Results and Analysis

For the legal domain, where many terms are exceptionally long compared to most other domains, our training scheme shows consistent improvements over the standard MLE counterparts, as shown in Table 3 and Table 4. Even with the extreme setting of law Ko→En, low-resourced and typologically divergent, our method is still effective in most metrics we use. Compared to the autoregressive models, GU19 and SUSANTO20 did not achieve competitive BLEU scores in our domain-specific setup. We suspect that this is due to both its complex decoding nature and the small amount of training data (originally WMT). Sampled translation results are reported in Table 9.

For the medical domain, the behaviors of two

baselines, VASWANI17 and CHEN20, are not clearly shown compared to the legal domain. However, our training scheme shows consistent improvements in BLEU and Term% at 2>micro which reflects the global performance of long terminology generation. Similar to the legal De→En results, SUSANTO20 shows better performance on several metrics on long terminology translation, but the BLEU score is decreased by about 8 points, compared to no dictionary use.

5 Conclusion

We propose a simple and effective training scheme for improving lexically constrained NMT by introducing the masked span prediction task on the decoder side. Our method shows its effectiveness in terms of terminology and sentence-level translation over the standard MLE training in highly specialized domains in two language pairs. As we publicly release our code and datasets, we hope that more people can join this area of research without much burden. In the future, we plan to further investigate applying our method to non-autoregressive methods.

Acknowledgments

We want to thank Minjoon Seo for his constructive comments on our research direction and Wooju Kim for his help in starting this topic. We also thank anonymous reviewers for their effort and valuable feedback. This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant (No.2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)), and National Research Foundation of Korea (NRF) grant (NRF-2020H1D3A2A03100945), funded by the Korea government (MSIT).

References

- Roei Aharoni and Yoav Goldberg. 2020. [Unsupervised domain clusters in pretrained language models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.
- Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof

- Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. [Findings of the 2020 conference on machine translation \(WMT20\)](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.
- Rajen Chatterjee, Matteo Negri, Marco Turchi, Marcello Federico, Lucia Specia, and Frédéric Blain. 2017. [Guiding neural machine translation decoding with external knowledge](#). In *Proceedings of the Second Conference on Machine Translation*, pages 157–168, Copenhagen, Denmark. Association for Computational Linguistics.
- Guanhua Chen, Yun Chen, Yong Wang, and Victor O.K. Li. 2020. [Lexical-constraint-aware neural machine translation via data augmentation](#). In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3587–3593. International Joint Conferences on Artificial Intelligence Organization. Main track.
- Chenhui Chu and Rui Wang. 2018. [A survey of domain adaptation for neural machine translation](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Josep Crego, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, Patrice Brunelle, Aurelien Coquad, Yongchao Deng, et al. 2016. Systran’s pure neural machine translation systems. *arXiv preprint arXiv:1610.05540*.
- 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.
- Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. [Training neural machine translation to apply terminology constraints](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.
- Tobias Domhan and Felix Hieber. 2017. [Using target-side monolingual data for neural machine translation through multi-task learning](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1500–1505, Copenhagen, Denmark. Association for Computational Linguistics.
- Miriam Exel, Bianka Buschbeck, Lauritz Brandt, and Simona Doneva. 2020. [Terminology-constrained neural machine translation at SAP](#). In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 271–280, Lisboa, Portugal. European Association for Machine Translation.
- Jiatao Gu, Changhan Wang, and Junbo Zhao. 2019. [Levenshtein transformer](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Hwei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.
- Dan Gusfield. 1997. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press.
- Eva Hasler, Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne. 2018. [Neural machine translation decoding with terminology constraints](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 506–512, New Orleans, Louisiana. Association for Computational Linguistics.
- Chris Hokamp and Qun Liu. 2017. [Lexically constrained decoding for sequence generation using grid beam search](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.
- Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime Carbonell. 2019. [Domain adaptation of neural machine translation by lexicon induction](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2989–3001, Florence, Italy. Association for Computational Linguistics.
- Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. [Google’s multilingual neural machine translation system: Enabling zero-shot translation](#). *Transactions of the Association for Computational Linguistics*, 5:339–351.
- Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2019. [SpanBERT: Improving pre-training by representing and predicting spans](#). *arXiv preprint arXiv:1907.10529*.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. [Moses: Open](#)

- source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.
- Philipp Koehn and Rebecca Knowles. 2017. **Six challenges for neural machine translation**. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.
- 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.
- Mathias Müller, Annette Rios, and Rico Sennrich. 2020. **Domain robustness in neural machine translation**. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 151–164, Virtual. Association for Machine Translation in the Americas.
- Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. **fairseq: A fast, extensible toolkit for sequence modeling**. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.
- Eunjeong L Park and Sungzoon Cho. 2014. **Konlpy: Korean natural language processing in python**. In *Annual Conference on Human and Language Technology*, pages 133–136. Human and Language Technology.
- Wei Peng, Chongxuan Huang, Tianhao Li, Yun Chen, and Qun Liu. 2020. **Dictionary-based data augmentation for cross-domain neural machine translation**. *arXiv preprint arXiv:2004.02577*.
- 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.
- Matt Post and David Vilar. 2018. **Fast lexically constrained decoding with dynamic beam allocation for neural machine translation**. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. **Exploring the limits of transfer learning with a unified text-to-text transformer**. *Journal of Machine Learning Research*, 21(140):1–67.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. **Improving neural machine translation models with monolingual data**. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. **Neural machine translation of rare words with subword units**. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.
- Kai Song, Kun Wang, Heng Yu, Yue Zhang, Zhongqiang Huang, Weihua Luo, Xiangyu Duan, and Min Zhang. 2020. **Alignment-enhanced transformer for constraining nmt with pre-specified translations**. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8886–8893.
- Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019a. **Code-switching for enhancing NMT with pre-specified translation**. 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 449–459, Minneapolis, Minnesota. Association for Computational Linguistics.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tiejun Liu. 2019b. **Mass: Masked sequence to sequence pre-training for language generation**. In *International Conference on Machine Learning*, pages 5926–5936. PMLR.
- Raymond Hendy Susanto, Shamil Chollampatt, and Liling Tan. 2020. **Lexically constrained neural machine translation with Levenshtein transformer**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3536–3543, Online. Association for Computational Linguistics.
- Brian Thompson, Rebecca Knowles, Xuan Zhang, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. **HABLex: Human annotated bilingual lexicons for experiments in machine translation**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1382–1387, Hong Kong, China. Association for Computational Linguistics.

- Jörg Tiedemann. 2012. [Parallel data, tools and interfaces in OPUS](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).
- 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 *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.
- Jiajun Zhang and Chengqing Zong. 2016a. Bridging neural machine translation and bilingual dictionaries. *arXiv preprint arXiv:1610.07272*.
- Jiajun Zhang and Chengqing Zong. 2016b. [Exploiting source-side monolingual data in neural machine translation](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Austin, Texas. Association for Computational Linguistics.

A Preprocessing and Training

A.1 Preprocessing

For De→En, we applied Moses tokenization (Koehn et al., 2007) and joint source-target byte pair encoding (BPE) (Sennrich et al., 2016b) with 20,000 split operations. For En→Ko, English was tokenized using spaCy⁶ and Korean using KoNLPy’s MeCab-ko⁷ (Park and Cho, 2014), followed by BPE with 20,000 operations. We apply sentence filtering up to 80 tokens.

A.2 Training

Model We follow the base Transformer architecture and fix the same hyperparameter configurations for all baselines. For the exact implementation of each baseline, we followed the authors’ official code on github (CHEN20⁸ and SUSANTO20⁹). We implemented our code using FAIRSEQ¹⁰ (Ott et al., 2019), trained on Nvidia GeForce RTX 3090 and RTX 2080 Ti GPUs.

Hyperparameter Detailed hyperparameter settings of baselines are reported below. Span masking and task coefficient only apply to our proposed training scheme.

Transformer	
Embedding dim.	512
Transformer FFN dim.	2048
Enc/Decoder layers	6
Attention heads	8
Share all embedding	True
Dropout	0.3
Label smoothing	0.1
Optimizer	Adam
Learning rate	0.0005
Warmup updates	4000
Maximum token per batch	4096
Maximum token lengths	80
Span Masking	
Span length	Geometric (p=0.2)
Maximum span length	10
Minimum span length	1
Corruption ratio	0.5
Task Coefficient	
Task coefficient (γ)	0.5

Table 6: Hyperparameter settings

⁶<https://spacy.io/>

⁷<https://konlpy.org/en/latest/>

⁸<https://github.com/ghchen18/leca>

⁹<https://github.com/raymondhs/constrained-levt>

¹⁰<https://github.com/pytorch/fairseq>

B Terminology-Aware Data Split Algorithm

Algorithm 1: Terminology-Aware Data Split Algorithm

Data: Dictionary \mathbb{D} , Corpus \mathbb{C} , Held-out Size R

Result: Sent=(Sent_{train}, Sent_{valid}, Sent_{test})
Term=(Term_{train}, Term_{valid}, Term_{test})

```

1: Sort  $\mathbb{D}$  in a descending order
2:  $N = \text{dict}()$ 
3:  $T' = \text{dict}()$ 
4:  $S = (\lvert\mathbb{C}\rvert - 2 * R, R, R)$ 
5: Senttrain = [], Sentvalid = [], Senttest = []
6: Termtrain = [], Termvalid = [], Termtest = []
7: for  $i$  in  $\{1, 2, \dots, \lvert\mathbb{C}\rvert\}$  do
8:    $x, y = \mathbb{C}[i]$ 
9:    $ngramlist = []$ 
10:   $T'' = []$ 
11:  for  $(x', y')$  in  $\mathbb{D}$  do
12:    if  $y'$  in  $y$  and  $x'$  in  $x$  then
13:       $ngramlist.append(ngram(y'))$ 
14:       $y = y.replace(y', "", 1)$ 
15:       $x = x.replace(x', "", 1)$ 
16:       $T''.append((x', y'))$ 
17:    end if
18:  end for
19:   $n = \text{Max}(ngramlist)$ 
20:  if  $n$  is not in  $N.keys()$  then
21:     $N[n] = []$ 
22:  end if
23:   $N[n].append(i)$ 
24:   $T'[i] = T''$ 
25: end for
26:  $K = \text{Sort the keys in } N \text{ in a descending order.}$ 
27: for  $k$  in  $K$  do
28:    $i_{dk}, i_{uk} = \text{DuplicateCheck}(N[k])$ 
29:   (Sent, Term) += Distributor.Dup( $i_{dk}$ ,  $N[k]$ ,  $T'$ ,  $S$ )
30:   (Sent, Term) += Distributor.Uni( $i_{uk}$ ,  $N[k]$ ,  $T'$ ,  $S$ )
31: end for

```

Line 1 : Sort \mathbb{D} w.r.t. target language terms.

Line 2 : Initialize a dictionary for storing paired sentences. The keys are the longest n-gram lengths for each sentence w.r.t. the target language.

Line 3 : Initialize a dictionary for storing matched terms. The keys are the indices of a corresponding sentence.

Line 13 : ngram() returns the token length of a term. In our case, it is used for calculating the length of a target language token y' .

Line 14 : Replace y' with "" in y to avoid unwanted substring duplication (e.g., In case of having “public officer” and “officer” in a sentence, we would like to first match “public officer” instead of “office” when we have “public officer” in the dictionary. See Line 1).

Line 19 : Calculate the maximum length of n-grams in y .

Line 23 : Store the sentence index w.r.t. its longest length of n-grams.

Line 24 : Store the list of terms w.r.t. its sentence index.

Line 28 : DuplicateCheck() checks for duplication in the corpus and returns duplicate and non-duplicate indices. Note

that i_{dk} is a list of duplicate sentence indices, and S_{uk} is a list of unique sentence indices.

Line 29 : Distributor_Dup() first calculates the number of sentences and phrases to be distributed across train, valid, and test sets following the ratio in S , and then distributes sentences accordingly.

Line 30 : Distributor_Uni() distributes unique sentences and phrases alternatively between train, valid, and test sets.

C Removing duplicates

As the OPUS datasets contain duplicate sentences (Aharoni and Goldberg, 2020), we further evaluate each model with unseen, unique test samples only. Similar to Tables 7 and 8, our training scheme outperforms its MLE counterparts. The Ko-En law corpus does not contain any duplicate sentence, and therefore the results are equivalent to those in Tables 3 and 4.

Model	Law (De→En)						
	BLEU	Term% (†)				LSM-3 (†)	
		1-gram	2-gram	2>micro	2>macro	2>micro	2>macro
<i>Without dictionary</i>							
GU19	68.14	93.71	91.87	72.32	46.24	85.05	74.22
VASWANI17	72.86	95.30	93.36	78.68	54.99	88.40	78.69
+SSP	73.15	95.57	93.54	79.98	59.18	89.64	81.53
CHEN20	71.89	95.04	93.54	78.83	54.55	88.43	79.63
+SSP	72.93	95.46	94.14	79.74	56.63	89.19	80.88
<i>With dictionary</i>							
SUSANTO20	59.23	94.10	92.89	82.86	68.22	95.28	93.36
CHEN20	70.90	96.36	94.01	77.26	51.97	89.12	80.06
+SSP	72.70	96.85	95.68	83.65	58.33	93.30	83.40

Table 7: Results on the law domain dataset with no duplication in data (De→En).

Model	Medical (De→En)						
	BLEU	Term% (†)				LSM-3 (†)	
		1-gram	2-gram	2>micro	2>macro	2>micro	2>macro
<i>Without dictionary</i>							
GU19	54.27	89.93	84.35	67.83	50.44	78.18	70.33
VASWANI17	58.29	90.09	84.51	70.98	57.82	79.18	76.45
+SSP	59.19	90.43	85.50	71.51	49.20	80.38	70.10
CHEN20	58.27	90.30	84.02	70.38	64.02	79.28	80.14
+SSP	59.49	90.57	83.53	71.25	64.36	80.29	78.70
<i>With dictionary</i>							
SUSANTO20	45.60	88.77	89.95	86.86	68.22	94.83	93.86
CHEN20	58.30	90.81	89.79	79.83	67.62	86.59	85.05
+SSP	60.30	93.10	91.10	85.19	79.26	91.34	90.51

Table 8: Results on the medical domain dataset with no duplication in data (De→En).

D Examples

Table 9 shows translation results of the baselines and our method.

Source	dieses Vorbringen wurde zurückgewiesen , da die einschlägigen Bestimmungen der Grundverordnung sehr wohl mit dem WTO- Übereinkommen zur Durchführung des Artikels VI des Allgemeinen Zoll- und Handelsabkommens 1994 und dem Übereinkommen über Subventionen und Ausgleichsmaßnahmen vereinbar sind .
VASWAN17	this claim was rejected because the relevant provisions of the basic Regulation are very compatible with the 1994 WTO Agreement on Implementation of Article VI of the General Agreement on Tariffs and Trade and the 1994 Agreement on Subsidies and Countervailing Measures .
+SSP	this claim was rejected as the relevant provisions of the basic Regulation are indeed consistent with the WTO Agreement on Implementation of Article VI of the General Agreement on Tariffs and Trade 1994 and the Agreement on Subsidies and Countervailing Measures .
CHEN20	this claim was rejected as the relevant provisions of the basic Regulation are , however , in any event , compatible with the WTO Agreement on the implementation of Article VI of the General Agreement on Tariffs and Trade 1994 and with the Agreement on Subsidies and Countervailing Measures .
+SSP	this claim was rejected because it is true that the relevant provisions of the basic Regulation are consistent with the Agreement on Implementation of Article VI of the General Agreement on Tariffs and Trade 1994 and the Agreement on Subsidies and Countervailing Measures .
Reference	this claim was rejected on the grounds that the anti-circumvention provisions of the basic Regulation are not incompatible with the Agreement on Implementation of Article VI of the General Agreement on Tariffs and Trade 1994 and the Agreement on Subsidies and Countervailing Measures .
Terminology	{ Übereinkommen zur Durchführung des Artikels VI des Allgemeinen Zoll- und Handelsabkommens 1994 → Agreement on Implementation of Article VI of the General Agreement on Tariffs and Trade 1994 , Übereinkommen über Subventionen und Ausgleichsmaßnahmen → Agreement on Subsidies and Countervailing Measures }
Source	(3) Das Angebot zur vorzugsweisen Zeichnung sowie die Frist , innerhalb deren dieses Recht ausgeübt werden muß , sind Gegenstand einer Bekanntmachung in dem gemäß der Richtlinie 68 / 151 / EWG bestimmten einzelstaatlichen Amtsblatt .
VASWAN17	3 . the tender for subscription and the time limit within which that right must be exercised shall be published in the national gazette determined in accordance with Directive 68 / 151 / EEC .
+SSP	3 . the tender for a preference call and the time limit within which that right must be exercised shall be the subject of a notice in the national gazette designated in accordance with Directive 68 / 151 / EEC .
CHEN20	3 . the tender for preferred subscription and the time limit within which it must be exercised shall be the subject of a notice published in the national publication designated pursuant to Directive 68 / 151 / EEC .
+SSP	3 . tenders for preference drawing together with the time limit within which that right must be exercised shall be the subject of a notice in the national gazette designated in accordance with Directive 68 / 151 / EEC .
Reference	any offer of subscription on a pre-emptive basis and the period within which this right must be exercised shall be published in the national gazette appointed in accordance with Directive 68 / 151 / EEC .
Terminology	{ Angebot zur vorzugsweisen Zeichnung → offer of subscription on a pre-emptive basis , Amtsblatt → national gazette }
Source	(19) Nach der Rechtsprechung des Gerichtshofs sind einzelstaatliche Vorschriften betreffend die Fristen für die Rechtsverfolgung zulässig , sofern sie für derartige Klagen nicht ungünstiger sind als für gleichartige Klagen , die das innerstaatliche Recht betreffen , und sofern sie die Ausübung der durch das Gemeinschaftsrecht gewährten Rechte nicht praktisch unmöglich machen .
VASWAN17	(19) According to the case law of the Court of Justice , national rules concerning time limits for bringing actions may be allowed , provided that such actions are not less favourable than those relating to the like actions under national law and do not make it impossible in practice to exercise the rights conferred by Community law .
+SSP	(19) According to the case law of the Court of Justice , national rules concerning the time limits for prosecution are admissible , provided that they are not less favourable to such actions than those for similar actions under national law if they do not make it impossible to exercise the rights conferred by Community law in practice .
CHEN20	(19) According to the case law of the Court of Justice , national rules on the time limits for the exercise of jurisdiction may be allowed , provided that they are not less favourable for such actions than for the same actions covered by national law and do not make it impossible in practice to exercise the rights conferred by Community law .
+SSP	(19) The Court of Justice has case-law that national provisions relating to time limits for bringing actions may be accepted , provided that they are no less favourable in such actions than those relating to similar actions brought under national law , provided that they do not practically make it impossible for the exercise of rights conferred by Community law .
Reference	(19) According to the case-law of the Court of Justice , national rules relating to time limits for bringing actions are admissible provided that they are not less favourable than time limits for similar actions of a domestic nature and that they do not render the exercise of rights conferred by the Community law impossible in practice .
Terminology	{ Gerichtshof → Court of Justice , Gemeinschaftsrecht → Community law , zulässig → admissible , Rechtsprechung → case-law }

Table 9: Translation outputs of the models trained with or without our method.