Lingua Custodia's participation at the WMT 2021 Machine Translation using Terminologies shared task

Melissa Ailem, Jinghsu Liu and Raheel Qader
Lingua Custodia, France
{melissa.ailem,jingshu.liu,raheel.qader}@linguacustodia.com

Abstract

This paper describes Lingua Custodia's submission to the WMT21 shared task on machine translation using terminologies. We consider three directions, namely English to French, Russian, and Chinese. We rely on a Transformer-based architecture as a building block, and we explore a method which introduces two main changes to the standard procedure to handle terminologies. The first one consists in augmenting the training data in such a way as to encourage the model to learn a copy behavior when it encounters terminology constraint terms. The second change is constraint token masking, whose purpose is to ease copy behavior learning and to improve model generalization. Empirical results show that our method satisfies most terminology constraints while maintaining high translation quality.

1 Introduction

Neural-based architectures have become standard for Machine Translation (MT), they are efficient and offer state-of-the-art performance in many scenarios (Vaswani et al., 2017). However, these models often trained on very large corpora turn out to be less adequate in domains that require very careful use of terminology. For instance, consider the following English sentence from a biomedical corpus ***"now for the fever you can take a tachipirina sweet"*** . The term ***"tachipirina sweet"*** refers to ***"paracétamol"*** in French. Unfortunately, a generic English-French Neural MT (NMT) model would translate the above sentence as: ***"maintenant pour la fièvre tu peux prendre un tachipirina bonbon"***, where the term ***"tachipirina sweet"*** is translated ***"tachipirina bonbon"***.

The goal of the WMT21 shared task on machine translation using terminology constraints is to explore methods that can take into account terminology constraints, in order to improve MT models' accuracy and consistency on specific domains. In the literature there are two main families of methods to take into account specific terminologies. One family incorporates terminology constraints at inference (Post and Vilar, 2018; Susanto et al., 2020). Members of this category can guarantee strict enforcement of constraints, however this often comes at the cost of higher decoding time and decreased accuracy (Hokamp and Liu, 2017; Post and Vilar, 2018). The other family of method integrates terminologies at training time (Dinu et al., 2019; Ailem et al., 2021), and they have the benefit of not changing the NMT model as well as of not incurring additional computational overheads at inference time (Crego et al., 2016; Song et al., 2019; Dinu et al., 2019).

We participate in the following three directions: English to French, Russian, and Chinese, and the system we submit falls into the second family of method incorporating terminologies at training time. More precisely, we explore a variant of the models proposed in (Ailem et al., 2021), which we train for each language pair. Following this work, we first annotate our training data with the constraints using tags to distinguish constraints terms from other tokens in the sentences. Second, we further perform constraint-token masking, which improves model robustness/generalization as supported by our experiments.

The rest of the paper is organized as follows: section 2 reviews the details of our system, section 3 describes the training data selection, the development and test sets, as well as the terminologies used for each language pair, and section 4 presents the different experimental settings and results.

2 Method

Our objective is to encourage neural machine translation to satisfy lexical constraints. To this end, we rely on the approch proposed in (Ailem et al., 2021), which introduces two changes to the standard procedure, namely training data augmentation

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 799–803
November 10–11, 2021. ©2021 Association for Computational Linguistics

Source	since COVID-19 shows similarities to **SARS-CoV** and MERS-CoV , it is likely that their effect on pregnancy are similar .
Constraints	**SARS-CoV** → **SARS-CoV**
TADA	since COVID-19 shows similarities to <S> **SARS-CoV** <C> **SARS-CoV** </C> and MERS-CoV , it is likely that their effect on pregnancy are similar .
+MASK	since COVID-19 shows similarities to <S> **MASK** <C> **SARS-CoV** </C> and MERS-CoV , it is likely that their effect on pregnancy are similar .

Figure 1: Illustration of TrAining Data Augmentation (TADA) and MASK.

and token masking. In the following we describe these two operations, which are also depicted in Figures 1 and 2.

TrAining Data Augmentation (TADA). The purpose of this step is to encourage the NMT model to exhibit a copy behavior when it encounters constraint terms whose translation should be consistent with some terminology. This step, illustrated in Figures 1 and 2, consists in using tags to annotate our training data with the terminology constraints, i.e., indicate the constraints (if any) in a given source sentence. Note that in the literature, there are other variants that use additional information such as source factors (Dinu et al., 2019). We do not use such information, and we specify terminologies using tags only.

Token MASKing (MASK). We further consider masking the source part of the constraint – tokens in blue – as illustrated in Figure 1 last row. As suggested in (Ailem et al., 2021), this masking strategy provides a more general pattern for the model to learn to perform the copy operation every time it encounters the tag $< S >$ followed by the MASK token. Moreover, this can make the model more apt to support conflicting constraints, i.e., constraints sharing the same source part but which have different target parts. This may be useful in situation in which some tokens must be translated into different targets for some specific documents and contexts at test time.

3 Data

This section provides information and some statistics regarding the datasets for the three language pairs we consider.

Training Data Selection. We consider three language pairs, namely English to French, Russian, and Chinese. Since our method acts at training time, we first perform a training data selection in order to obtain a reasonable number of sentences containing at least one term from the provided terminologies. To do so, we consider both bilingual and monolingual data, provided as part of the shared task. In fact, we observe that bilingual data do not contain many sentences with terminology terms. Thus, we rely on back-translation of monolingual data, which contains more recent news on COVID-19, to obtain more sentence pairs with terminologies. We rely on OpusMT[1] to back translate the Russian monolingual to English. For Chinese and French we use in-house translation engines. Note that we further convert the Chinese data into simplified Chinese using OpenCC. Following previous work on terminology control (Dinu et al., 2019; Ailem et al., 2021), only 10% of the training sentences are annotated in order to maintain the model's performance in terminology free cases. The details about training data selection for the different language pairs are summarized in tables 1, 2 and 3.

Development and Test Sets. For all language pairs, a development and test sets are provided. Note that for the test sets we have access to the source part only. For the dev sets, the terminology constraints associated with each sentence are available, for the test sets this information is not available, and we leverage the terminology files to find constraint terms in these sets. Just like the training data, both test and dev sets are augmented with the terminology constraints as presented in figures 1 and 2. The dev/test sets of the different language pairs share the same English source file containing 971/2100 sentences respectively.

Terminologies. For each language pair, we use the provided terminologies to annotate our train, dev and test sets. The terminologies consist of respectively 670, 925 and 710 unique source-target terms for English → French, Russian and Chinese. We also observe that one source term might be associated with one or more target terms. In that

[1] https://github.com/Helsinki-NLP/Opus-MT

Source	the Canadian government announced CA $ 275 million in funding for 96 research projects on medical countermeasures against COVID-19 , including numerous **vaccine** candidates at Canadian universities , with plans to establish a " vaccine bank " of new vaccines for implementation if another Coronavirus outbreak occurs .
Constraints	**vaccine** → **vaccin**, **vaccines** → **vaccins**, **Coronavirus outbreak** → **épidémie de coronavirus**
TADA	the Canadian government announced CA $ 275 million in funding for 96 research projects on medical countermeasures against COVID-19 , including numerous <S> **vaccine** <C> **vaccin** </C> candidates at Canadian universities , with plans to establish a " <S> **vaccine** <C> **vaccin** </C> bank " of new <S> **vaccines** <C> **vaccins** </C> for implementation if another <S> **Coronavirus outbreak** <C> **épidémie de coronavirus** </C> occurs .
+MASK	the Canadian government announced CA $ 275 million in funding for 96 research projects on medical countermeasures against COVID-19 , including numerous <S> **MASK** <C> **vaccin** </C> candidates at Canadian universities , with plans to establish a " <S> **MASK** <C> **vaccin** </C> bank " of new <S> **MASK** <C> **vaccins** </C> for implementation if another <S> **MASK MASK** <C> **épidémie de coronavirus** </C> occurs .

Figure 2: Illustration of TrAining Data Augmentation (TADA) and MASK (multiple constraints in one sentence).

Data type	#sentences	#term-grounded sentences	Corpora
Monolingual fr	342,941	342,941	News Crawl 2020
Parallel en-fr	3,110,291	110,291	NCv16, UN, Common Crawl, Europarl v10
Parallel en-fr (biomedical)	1,733,757	67,887	EMEA, Medline Titles, Medline abstracts
#Total	5,186,989	521,119	

Table 1: English → French data we use for training.

Data type	#sentences	#term-grounded sentences	Corpora
Monolingual ru	997,889	697,889	News Commentary, News
Parallel en-ru	6,121,064	3,169	News Commentary, Wikititles, ParaCrawl, UN, Wikimatrix, Common Crawl, Yandex
Parallel en-ru (biomedical)	46,782	0	Medline
#Total	7,165,738	701,058	

Table 2: English → Russian data we use for training.

Data type	#sentences	#term-grounded sentences	Corpora
Monolingual zh	899,163	899,163	News Crawl 2020
Parallel en-zh (up-sampled)	12,900	12,900	Wikititles
Parallel en-zh	6,322,275	0	NCv16, ParaCrawl, Wikimatrix, UN, CCMT
#Total	7,234,338	912,063	

Table 3: English → Chinese data we use for training.

case, when annotating the train and dev sets we choose the target term used in the ground truth translation. For the test set, we select one of the possible terms at random.

4 Experimental results

4.1 Settings

For English to French and Russian pairs, we first tokenize the terminology files and the train/test/dev sets before annotating them with the terminology constrains. We use the Moses tokenizer (Koehn et al., 2007) for this step. We then rely on BPE encoding (Sennrich et al., 2015) with 40k merge operations to segment words into subword-units, which results in a joint vocabulary size of 42588 words for English->French, and vocabulary sizes of (44644, 47532) for the (English, Russian) pair. For English->Chinese we rely on sentence piece (Kudo and

Model	BLEU	Exact-Match Accuracy	Window Overlap (2)	Window Overlap (3)	1-TERm	COMET
Transformer	32.12	0.325	0.112	0.114	0.369	0.023
Constrained decoder	40.12	0.856	0.306	0.298	0.535	0.416
TAG+MASK	**44.90**	**0.919**	**0.344**	**0.335**	**0.598**	**0.681**

Table 4: Comparison of different models on the English → French test set.

Language Pair	BLEU	Exact-Match Accuracy	Window Overlap (2)	Window Overlap (3)	1-TERm	COMET
English → French	44.90	0.919	0.344	0.335	0.598	0.681
English → Russian	29.13	0.849	0.247	0.248	0.474	0.604
English → Chinese	29.16	0.829	0.223	0.225	0.437	0.637

Table 5: Results of the investigated system (TAG+MASK) across all the language pairs we consider. Results are obtained using the test set.

Richardson, 2018) for tokenization, which also performs BPE encoding simultaneously and results in a vocabulary size of 52172 for Chinese and 39996 for english. We then annotate the train/test/dev sets with the terminology constraints.

As a building block for our system, we use the transformer architecture (Vaswani et al., 2017) with 6 stacked encoders/decoders and 8 attention heads. For English-French, the source and target embeddings are tied with the softmax layer. We use 512-dimensional embeddings, 2048-dimensional inner layers for the fully connected feed-forward network and a dropout rate of 0.3. The models are trained for a minimum of 50 epochs and a maximum of 100 epochs with a batch size of 2000 tokens per iteration and an initial learning rate of 5×10^{-4}. For each language pair, the validation set is used to compute the stopping criterion. We use a beam size of 5 during inference for all models.

4.2 Results

For all language pairs, the models are evaluated using the standard MT evaluation metrics (BLEU and COMET scores) as well as other terminology-targeted metrics (Anastasopoulos et al., 2021). The latter include the "Exact-Match Accuracy" measure, which simply compute the percentage of constraint terms present in the predicted translations. Although this measure provides an indication of terminology satisfaction, it can only assess whether a term is present in the hypotheses without evaluating whether this target term is correctly placed. To overcome this issue, the authors in (Anastasopoulos et al., 2021) proposed an additional measure, namely "Window Overlap", which computes the percentage of similar tokens surrounding the constraint terms – within a defined window – in the ground truth and the generated hypotheses. Finally, the models are also evaluated in terms of "Terminology-biased TER" score, which is an edit distance based metric (Snover et al., 2006; Anastasopoulos et al., 2021).

We compare the our model TAG+MASK with the traditional transformer baseline (Vaswani et al., 2017) and the constrained decoder approach (Post and Vilar, 2018), which integrates the constraints during inference time. Results on English → French data are presented in table 4. We observe that the TAG+MASK approach significantly improves over baselines in terms of all measures.

Table 5 depicts the results that the submitted system reaches across all the language pairs in terms of different metrics.

5 Conclusion

In this paper, we describe our submission to the WMT21 shared task on machine translation using terminologies. We participate in three language pairs, namely English → French, Russian and Chinese. Our system integrates terminology constraints during training by augmenting the data with terminological terms. Due to the lack of parallel training data containing the terminology terms, we rely on monolingual data for all language pairs to augment the number of sentences containing terminology terms. Empirical results comparing our approach with terminology grounded as well as terminology free baselines show the effectiveness of the investigated method.

Acknowledgments

This work was partially funded by the French Ministry of Defense.

References

Melissa Ailem, Jingshu Liu, and Raheel Qader. 2021. Encouraging neural machine translation to satisfy terminology constraints. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1450–1455, Online. Association for Computational Linguistics.

Antonios Anastasopoulos, Laurent Besacier, James Cross, Matthias Gallé, Philipp Koehn, Vassilina Nikoulina, et al. 2021. On the evaluation of machine translation for terminology consistency. *arXiv preprint arXiv:2106.11891*.

Josep Crego, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, Patrice Brunelle, Aurelien Coquard, Yongchao Deng, et al. 2016. Systran's pure neural machine translation systems. *arXiv preprint arXiv:1610.05540*.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, page 3063–3068.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, page 1535–1546.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. *Proceedings of NAACL-HLT 2018*, page 1314–1324.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, page 1715–1725.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019. Code-switching for enhancing nmt with pre-specified translation. *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)*, page 449–459.

Raymond Hendy Susanto, Shamil Chollampatt, and Liling Tan. 2020. Lexically constrained neural machine translation with levenshtein transformer. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 3536–3543.

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*, pages 5998–6008.