Lingua Custodia's participation at the WMT 2022 Word-Level Auto-completion shared task

Melissa Ailem, Jinghsu Liu, Jean-Gabriel Barthélemy and Raheel Qader
Lingua Custodia, France
{melissa.ailem,jingshu.liu,j-g.barthelemy,raheel.qader}@linguacustodia.com

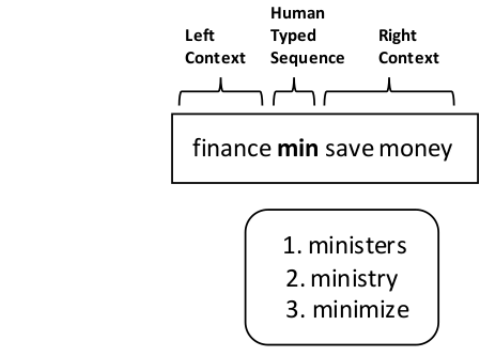
Abstract

This paper presents Lingua Custodia's submission to the WMT22 shared task on Word Level Auto-completion (WLAC). We consider two directions, namely German-English and English-German. The WLAC task in Neural Machine Translation (NMT) consists in predicting a target word given few human typed characters, the source sentence to translate, as well as some translation context. Inspired by recent work in terminology control, we propose to treat the human typed sequence as a constraint to predict the right word starting by the latter. To do so, the source side of the training data is augmented with both the constraints and the translation context. In addition, following new advances in WLAC, we use a joint optimization strategy taking into account several types of translation context. The automatic as well as human accuracy obtained with the submitted systems show the effectiveness of the proposed method.

1 Introduction

Modern advances in Neural Machine Translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2014; Luong et al., 2015; Vaswani et al., 2017) gave rise to a new era, where the translation quality significantly surpasses previous statistical machine translation (SMT) models (Och and Ney, 2002; Koehn et al., 2003; Koehn, 2010).

Although these approaches generate high quality translations, there is still a long way to go towards meeting human quality. In fact, NMT models can still generate several types of grammatical and/or semantic mistakes, which is not tolerated in scenarios requiring accurate and prompt translations. These scenarios include for instance the translations of legal and financial documents, where mistakes are not permitted and can be costly. To overcome this issue, several Computer-aided translation (CAT) systems have been proposed (Knowles and Koehn, 2016; Santy et al., 2019) to refine NMT models. CAT tools include for instance Automatic Post-edition (Junczys-Dowmunt and Grundkiewicz, 2017; Correia and Martins, 2019; Lopes et al., 2019), terminology control (Hokamp and Liu, 2017; Post and Vilar, 2018; Dinu et al., 2019; Ailem et al., 2021) and sentence vs word-level auto-completion (Knowles and Koehn, 2016; Zhao et al., 2020; Li et al., 2021).



Source Die Finanzminister sind der Meinung , man müsse sparen .

Reference The finance ministers believe it is important to save money .

Figure 1: Given the source sentence to translate, the translation contexts, and the human typed characters, the WLAC task aims to predict a target word starting by the human typed sequence. As illustrated, the word to predict is not necessarily consecutive to the left and right contexts.

The current shared task is on Word-Level Auto-Completion (WLAC) methods, whose objective, as illustrated in Figure 1, is to predict a target word given a source sentence, a translation context, and at least one human typed character. WLAC is a central Computer-aided task as it helps human translators generate diverse translations quickly and effectively. Unfortunately, due to the lack of benchmark datasets, very little work has considered this task. Existing methods include the work of Huang et al. (2015), where the authors leverage the source sen-

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 1170–1175
Abu Dhabi, December 7–8, 2022.

tence as well as human typed characters to predict the target word. More recently, Li et al. (2021) proposed to use context information in addition to human-typed characters and source sentence. Furthermore, the authors presented a generic procedure to simulate WLAC data from any parallel translation datasets, and proposed the first public benchmark for this task. The benchmark dataset contains several types of contexts and therefore a joint optimization strategy is used to take into account all context types during training.

We participate in two directions, namely English-German (EN-DE), German-English (DE-EN), and we submitted four systems, two for each language direction. Following previous work (Li et al., 2021), our method leverages source sentence, translation context as well as human-typed characters, and it uses a joint objective function to learn model parameters on different types of contexts simultaneously. Furthermore, inspired by recent progress in Terminology Control (TC) for NMT (Dinu et al., 2019; Ailem et al., 2021), we propose a new WLAC method that treats the human typed sequence as a constraint to generate the right word. To do so, we augment our training data with translation context as well as human typed characters (constraints). We use tags where needed to distinguish these terms from source tokens.

The rest of the paper is organized as follows. Section 2 describes the details of our system, section 3 presents the data, while section 4 shows the different experimental settings and results.

2 Method

Herein we present our WLAC approach which is inspired by recent advances in this task (Li et al., 2021) as well as recent work on terminology control (Ailem et al., 2021).

2.1 Data Annotation

Inspired by previous work on Terminology Control (Ailem et al., 2021), the idea here is to consider human typed characters as a constraint. In particular, the objective is to constrain the NMT model to predict a word that, obligatorily, starts with human typed characters. To do so, we augment the source side of our training data with the translation context as well as the human typed sequence of characters. Furthermore, we use tags to specify the constraints (human typed characters) in the context translation where relevant, and use the special token MASK in order to provide a more general pattern for the model to learn how to predict the right word starting with human sequence. The WLAC data provided by the WMT task and used in (Li et al., 2021) contains 4 types of context, namely left and right contexts (bi-context), left context only (prefix), right context only (suffix), and no context at all (zero context). The different annotations according to each context types are depicted in table 1.

2.2 Joint Cross-Entropy Loss

Let $\mathbf{x} = (x_1, x_2, ..., x_m)$ denotes the input sentence to translate, $\mathbf{s} = (s_1, s_2, \ldots, s_k)$ a sequence of human typed characters, and $\mathbf{c} = (c_l, c_r)$ the translation context, where $c_l = (c_{l,1}, c_{l,2}, \ldots, c_{l,i})$ denotes the left context, while $c_r = (c_{r,1}, c_{r,2}, \ldots, c_{r,j})$ denotes the right context. The objective of the WLAC task is to predict a word w given a source sequence $\mathbf{x}$, human typed sequence $\mathbf{s}$ and a translation context $\mathbf{c}$ in order to establish a partial translation. The training data $\mathcal{D}$ of a WLAC task can be described as a set of tuples $(\mathbf{x}, \mathbf{s}, \mathbf{c}, w)$. From a probabilistic perspective, a WLAC task can be cast as estimating the conditional distribution $p(w|\mathbf{x}, \mathbf{c}, \mathbf{s})$. Since there is different types of context (as described in section 2.1), we follow the work of Li et al. (2021) and adopt a joint training strategy. In particular, the four types of context are considered during training giving rise to the following loss function:

$$
\begin{aligned}
\mathcal{L} &= -\log p(w|\mathbf{x}, \mathbf{c}, \mathbf{s}) \\
&= -\sum_{(\mathbf{x},\mathbf{c},\mathbf{s},w)\in\mathcal{D}_{bi}} \log p(w|\mathbf{x}, c_l, c_r, \mathbf{s}) \\
&- \sum_{(\mathbf{x},c_r,\mathbf{s},w)\in\mathcal{D}_{suf}} \log p(w|\mathbf{x}, c_r, \mathbf{s}) \\
&- \sum_{(\mathbf{x},c_l,\mathbf{s},w)\in\mathcal{D}_{pre}} \log p(w|\mathbf{x}, c_l, \mathbf{s}) \\
&- \sum_{(\mathbf{x},\mathbf{s},w)\in\mathcal{D}_{zero}} \log p(w|\mathbf{x}, \mathbf{s})
\end{aligned}
\tag{1}
$$

where $\mathcal{D}_{bi}$, $\mathcal{D}_{suf}$, $\mathcal{D}_{pre}$, $\mathcal{D}_{zero}$ correspond respectively to bi-context, suffix context, prefix context and zero context.

3 Data

We participate in two directions, namely English-German and German-English. We use the parallel

Source	Seebarsch gebacken auf seinem Rücken , fein zerschnippelter Porree und Zitronenmelissekraut .			
Target	Bar baked on the back with finely chopped leek and lemon melissa herbs .			
WLAC training data				
Input	bi-context	Seebarsch gebacken auf seinem Rücken , fein zerschnippelter Porree und Zitronenmelissekraut . <SEP> Bar baked on <S> MASK <C> wit </C> and lemon	Output	with
	Prefix Context	Seebarsch gebacken auf seinem Rücken , fein zerschnippelter Porree und Zitronenmelissekraut . <SEP> Bar baked on the back with finely <S> MASK <C> chop </C>		chopped
	Suffix Context	Seebarsch gebacken auf seinem Rücken , fein zerschnippelter Porree und Zitronenmelissekraut . <SEP> <S> MASK <C> B </C> lemon melissa herbs .		Bar
	Zero Context	Seebarsch gebacken auf seinem Rücken , fein zerschnippelter Porree und Zitronenmelissekraut . <SEP> <S> MASK <C> bak </C>		baked

Table 1: Illustration of our German-English training data. The WLAC training data can be build from traditional parallel translation data. During sampling, for each parallel sentences four samples are generated corresponding to four types of translation context, namely left and right contexts (bi-context), left context (prefix), right context (suffix) and no context at all (zero). The source side of the training data is a concatenation of the German source side and the English translation context separated by the tag **<SEP>**. Translation context is also augmented with human typed characters, which are considered as a constraint to orient the model to predict the right word. The tags <S>, <C> and </C> are added to differentiate between the constraints and other tokens in the input.

English-German data provided by the WLAC task consisting of almost 4.5M parallel sentences. Following task instructions, we use the script proposed in (Li et al., 2021) to simulate the WLAC training data from the provided classical translation data.

3.1 Parallel Data Cleaning

Before creating the WLAC samples, we apply several cleaning steps on the data to eliminate bad alignments. First, the data is re-segmented using the Python package pySBD (Sadvilkar and Neumann, 2020) in order to detect sentence boundaries. This step increases the number of parallel sentences to almost 6.5 M. Second, each parallel entry is scored between 0 and 1 using several tools. These tools include bicleaner (Ramírez-Sánchez et al., 2020), similarity scoring using LaBSE (Feng et al., 2020), and bicleaner-ai, which is inspired by the BERT-based model proposed in (Açarçiçek et al., 2020) for sentence classification. Table 2 presents the different scoring thresholds used to clean the parallel corpus. In particular, we rely on a combination of bi-cleaner and similarity scoring as well as bicleaner-ai. In our experiments, we consider both initial uncleaned data (Noisy) and the cleaned data. In addition to these scoring, we also rely on fasttext (Bojanowski et al., 2017) to eliminate sentence pairs identified as written in the wrong language (e.g., A french sentence in an English-German parallel corpus). After the cleaning we obtain a corpus of around 2.7 M parallel segments.

	Clean	Noisy
Bicleaner + Similarity	>1.4	>0
Bicleaner-ai	>0.25	>0.1
Total sentences	2 717 737	4 404 427

Table 2: The different thresholds applied on the corpus. Noisy corresponds to the original parallel data provided by the WLAC task. The threshold 1.4 is a combination of Bi-cleaner and Similarity scoring thresholds.

3.2 WLAC Data Construction

The parallel data commonly used for NMT and provided by the WLAC task cannot be used directly to train a WLAC model. Thus, following the task instructions, we use the script proposed in (Li et al., 2021) to simulate several samples for the WLAC training. For each sentence pair, 4 samples are created according to the four context types as presented in table 1. Since the provided initial data contains almost 4.5M parallel sentences, we obtain a WLAC corpus of almost 18M entries (4.5×4). As presented in the previous section, we have also used a cleaned version of the provided corpus, containing around 2.7M entries. For the

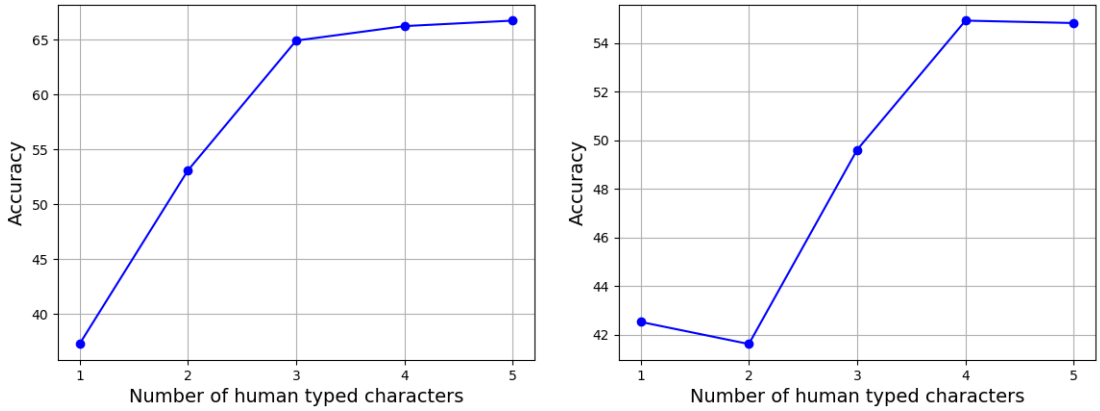


Figure 2: Accuracy obtained with different number of human typed characters. Left : German-English system with initial corpus. Right : English-German system with initial corpus.

latter, we obtain around 10.8M WLAC training samples. Hence, synthetic WLAC training data are build for the two corpus versions (clean and initial) in the two considered directions: English-German and German-English. The dev sets are build from 3000 EN-DE and DE-EN parallel sentences from the initial corpus. To do so, the same sampling script is used resulting in 20K entries for both directions. The test sets released by the WLAC task contain 29596 and 25895 samples for DE-EN and EN-DE respectively.

4 Experiments

4.1 Settings

We use a Transformer architecture (Vaswani et al., 2017) with 6 stacked encoders/decoders and 8 attention heads as a building block for our systems. For both EN-DE and DE-EN, 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 the validation set is used to compute the stopping criterion[1]. We use a batch size of 4000 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.

Before annotating our corpus as presented in table 1, we first tokenize the data using Moses tokenizer (Koehn et al., 2007). After augmenting the data with translation context and human typed sequence, we perform a BPE encoding (Sennrich et al., 2015) with 40k merge operations to segment words into subword-units, which results in a joint vocabulary size of around 44K tokens for both German-English and English-German.

Accuracy (%)		
	German-English	English-German
Cleaned Corpus	54.84	48.43
Initial Corpus	57.36	48.97
Human Evaluation (%)		
Cleaned Corpus	74.50	61.00
Initial Corpus	76.75	61.75

Table 3: Accuracy and Human Evaluation results.

4.2 Results

For both considered directions, the systems are evaluated using the Accuracy measure, corresponding to the percentage of correctly predicted words. This automatic accuracy is obtained using one single ground truth word for each sample. However, one sentence may have multiple translations, thus several Ground Truth are possible, making the automatic accuracy inadequate. To overcome this limitation, a human evaluation is applied on 400 randomly sampled entries from the test set. In particular, given the human typed sequence, the translation context and the source sentence to trans-

[1]The stopping criterion corresponds to 5 successive epochs without decreasing the validation loss function.

late, human annotators judge whether a predicted word can be correct according to the given context. The results obtained with our systems are presented in table 3.

Surprisingly, we observe that cleaning the different corpora is mirrored by a deterioration in results. Indeed, the best results are reached with the systems using initial training corpus. This might be due to the excessive cleaning, removing some scenarios that could be present in the test set.

Furthermore, we notice that the chances of predicting the right word are positively related with the number of human typed characters. We present in figure 2 the accuracy obtained with different numbers of human typed characters. In both directions, we observe that the accuracy improves with the typed sequence length. This is natural, as with few typed characters, several choices are possible, especially when the translation context is restricted or even non-existent (zero context situation).

5 Conclusion

This paper describes our submission to the WLAC shared task. We participate in two language directions, EN-DE and DE-EN, and submitted two systems for each direction. For each direction, the first system is trained using initial data provided by the task, while the second system is trained on cleaned data. The evaluation in terms of accuracy shows the effectiveness of the proposed method. Furthermore, a significant improvement of accuracy is observed when the number of human typed characters is greater than 1, suggesting that entering at least two characters restrain the search space and improve the chances to predict the right word.

References

Haluk Açarçiçek, Talha Çolakoğlu, Pınar Ece Aktan Hatipoğlu, Chong Hsuan Huang, and Wei Peng. 2020. Filtering noisy parallel corpus using transformers with proxy task learning. In *Proceedings of the Fifth Conference on Machine Translation*, pages 940–946, Online. Association for Computational Linguistics.

Melissa Ailem, Jinghsu Liu, and Raheel Qader. 2021. Encouraging neural machine translation to satisfy terminology constraints. *arXiv preprint arXiv:2106.03730*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Gonçalo M Correia and André FT Martins. 2019. A simple and effective approach to automatic post-editing with transfer learning. *arXiv preprint arXiv:1906.06253*.

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.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding.

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.

Guoping Huang, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2015. A new input method for human translators: integrating machine translation effectively and imperceptibly. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2017. An exploration of neural sequence-to-sequence architectures for automatic post-editing. *arXiv preprint arXiv:1706.04138*.

Rebecca Knowles and Philipp Koehn. 2016. Neural interactive translation prediction. In *AMTA (1)*, pages 107–120.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press.

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.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistic*, pages 127–133, Edmondton, Canada.

Huayang Li, Lemao Liu, Guoping Huang, and Shuming Shi. 2021. Gwlan: General word-level autocompletion for computer-aided translation. *arXiv preprint arXiv:2105.14913*.

António V Lopes, M Amin Farajian, Gonçalo M Correia, Jonay Trénous, and André FT Martins. 2019. Unbabel's submission to the wmt2019 ape shared task: Bert-based encoder-decoder for automatic post-editing. *arXiv preprint arXiv:1905.13068*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, page 1412–1421.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 295–302, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz-Rojas. 2020. Bifixer and bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal. European Association for Machine Translation.

Nipun Sadvilkar and Mark Neumann. 2020. PySBD: Pragmatic sentence boundary disambiguation. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 110–114, Online. Association for Computational Linguistics.

Sebastin Santy, Sandipan Dandapat, Monojit Choudhury, and Kalika Bali. 2019. Inmt: Interactive neural machine translation prediction. 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): System Demonstrations*, pages 103–108.

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.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*.

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.

Tianxiang Zhao, Lemao Liu, Guoping Huang, Huayang Li, Yingling Liu, Liu GuiQuan, and Shuming Shi. 2020. Balancing quality and human involvement: An effective approach to interactive neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9660–9667.