

Rule-based Morphological Inflection Improves Neural Terminology Translation

Weijia Xu

University of Maryland
weijia@cs.umd.edu

Marine Carpuat

University of Maryland
marine@cs.umd.edu

Abstract

Current approaches to incorporating terminology constraints in machine translation (MT) typically assume that the constraint terms are provided in their correct morphological forms. This limits their application to real-world scenarios where constraint terms are provided as lemmas. In this paper, we introduce a modular framework for incorporating lemma constraints in neural MT (NMT) in which linguistic knowledge and diverse types of NMT models can be flexibly applied. It is based on a novel cross-lingual inflection module that inflects the target lemma constraints based on the source context. We explore linguistically motivated rule-based and data-driven neural-based inflection modules and design English-German health and English-Lithuanian news test suites to evaluate them in domain adaptation and low-resource MT settings. Results show that our rule-based inflection module helps NMT models incorporate lemma constraints more accurately than a neural module and outperforms the existing end-to-end approach with lower training costs.¹

1 Introduction

Incorporating terminology constraints in machine translation (MT) has proven useful to adapt translation lexical choice to new domains (Hokamp and Liu, 2017) and to improve its consistency in a document (Ture et al., 2012). In neural MT (NMT), most prior work focuses on incorporating terms in the output exactly as given, using soft (Song et al., 2019; Dinu et al., 2019; Xu and Carpuat, 2021) or hard constraints (Hokamp and Liu, 2017; Post and Vilar, 2018). These approaches are problematic when translating into morphologically rich languages where terminology should be adequately inflected in the output, while it is

more natural and flexible to provide constraints as lemmas as in a dictionary.

To the best of our knowledge, only one paper has directly addressed this problem for neural MT: (Bergmanis and Pinnis, 2021) design an NMT model trained to copy-and-inflect the terminology constraints using target lemma annotations (TLA) — TLA are synthetic training samples where the source sentence is tagged with automatically generated lemma constraints. While this approach improves translation quality, the end-to-end training set-up prevents fast adaptation to lemmas and inflected forms that are rare or unseen at training time. Its impact is also limited to a specific neural architecture, and it is unclear whether its benefits port to more generic sequence-to-sequence models.

In this paper, we introduce a modular framework for inflecting terminology constraints in NMT. It relies on a cross-lingual inflection module that predicts the inflected form of each lemma constraint based on the source context only. The inflected lemmas can then be incorporated into NMT using any of the aforementioned constrained NMT techniques. Compared with TLA, this framework is more flexible, as it can be applied to diverse types of NMT architectures and inflection modules, and facilitates fast adaptation to new terminologies without retraining the base NMT model from scratch. This flexibility is enabled by the cross-lingual nature of the inflection module, which predicts the inflected form of each target lemma based on the source context only. This differs from traditional inflection models that predict the inflected forms based on pre-specified morphological tags or monolingual target context.

Based on this framework, this paper makes the following contributions:

- We construct and release test suites to evaluate models' ability to inflect terminology constraints for domain adaptation (English-German Health) and low-resource

¹Code and test suites are released at <https://github.com/Izecson/terminology-translation>

MT (English-Lithuanian News).

- We show that integrating linguistic knowledge through a simple rule-based inflection module improves over its neural counterpart in intrinsic and end-to-end MT evaluations.
- Our framework improves autoregressive and non-autoregressive translation, and outperforms the existing TLA approach for inflecting terminology translation. We open-source the code to facilitate replication and extensions.

2 Background

Autoregressive NMT with Constraints Terminology constraints can be incorporated in autoregressive NMT models via 1) constrained decoding where constraint terms are incorporated in the beam search algorithm (Hokamp and Liu, 2017; Post and Vilar, 2018), or 2) constrained training where NMT models are trained to incorporate constraints using synthetic parallel data augmented with constraint terms on the source side (Song et al., 2019; Dinu et al., 2019). These approaches all assume that the constraints are provided in the correct inflected forms and can be directly copied to the target sentence. Bergmanis and Pinnis (2021) extended the constrained training approach of Dinu et al. (2019) to incorporate lemma-form constraints in an end-to-end way – the inflected form of the lemma constraints are predicted jointly during translation. This approach requires a dedicated NMT model architecture to integrate constraints as additional inputs to the encoder, and learns inflection solely from the parallel data. By contrast, our approach can be applied to multiple NMT architectures and uses linguistically motivated rule that generalize better to rare and unseen terms.

Non-Autoregressive NMT with Constraints Instead of generating the output sequence incrementally from left to right, non-autoregressive NMT generates tokens in parallel (Gu et al., 2018; van den Oord et al., 2018; Ma et al., 2019) or by iteratively editing an initial sequence (Lee et al., 2018; Ghazvininejad et al., 2019). Architectures differ with the nature of edit operations: the Levenshtein Transformer (Gu et al., 2019) relies on insertion and deletion, while EDITOR (Xu and Carpuat, 2021) uses insertion and reposition (where each input token can be repositioned or deleted). Edit-based non-autoregressive generation provides a natural way to incorporate constraints in NMT – the constraints

can be put into the initial sequence and edited to produce the final translation (Susanto et al., 2020; Xu and Carpuat, 2021; Wan et al., 2020). Our approach can augment this family of techniques by inflecting constraints before they are used for further editing.

Morphological Inflection Morphological inflection is the process of alternating the morphological form of a lexeme that adds morpho-syntactic information of the word in a sentence (e.g. tense, case, number). Traditionally, morphological inflection as computational task is framed as predicting the inflected form of a word given its lemma and a set of morphological tags (e.g. N;ACC;PL represents a plural noun used in accusative case) (Cotterell et al., 2017). The task was traditionally tackled using hand-engineered finite state transducer that relies on linguistic knowledge (Koskenniemi, 1984; Kaplan and Kay, 1994), while recent work has shown impressive results by modeling it using neural sequence-to-sequence models (Faruqui et al., 2016). More recently, a context-based inflection task has been proposed where the inflected form of a lemma is predicted given the rest of the sentence as context (Cotterell et al., 2018). The state-of-the-art models for the task are neural models trained on supervised data (Cotterell et al., 2018; Kementchedjheva et al., 2018). The inflection module in our framework differs from those for the context-based inflection task in that it requires cross-lingual context-based inflection – it predicts the inflected form of a target lemma based only on the source language context.

Morphologically-Aware Translation In phrase-based MT, modeling morphological compounds on the source (Koehn and Knight, 2003) and target sides (Cap et al., 2014) improves translation quality. In NMT, morphologically-aware segmentation is also useful when translating from or into morphologically complex languages (Huck et al., 2017; Ataman and Federico, 2018; Banerjee and Bhat-tacharyya, 2018). Tamchyna et al. (2017) propose to overcome data sparsity caused by inflection by training NMT models to predict the lemma form and morphological tag of each target word. Different from prior work, we incorporate grammatical and morphological knowledge in an inflection module for terminology constraints in NMT.

3 Inflecting Target Lemmas Given the Source Context

We introduce a modular framework for inflecting terminology constraints for NMT, where we first build an inflection module that predicts the inflected form of each target lemma term based on the source sentence and then incorporate the inflected constraints in NMT using any of the aforementioned techniques. By framing the problem this way, we assume that the inflected forms can be inferred based only on the source context and integrated in a fluent translation by NMT models. In cases where there are multiple possible inflected forms corresponding to different ways of translating the source, the inflection module can predict one of the possible forms, and the NMT model can generate a translation conditioned on the predicted forms of the constraints. Compared with [Bergmanis and Pinnis \(2021\)](#), our framework is more flexible – it can be combined with any NMT model that enables translation with constraints and can leverage diverse types of morphological inflection modules in which linguistic knowledge can be easily incorporated.

Formally, given a source sequence \mathbf{x} and k target lemma words $\bar{\mathbf{z}} = (\bar{z}_1, \bar{z}_2, \dots, \bar{z}_k)$ that need to be inflected, the inflection module Θ predicts the inflected form of each target lemma $\mathbf{z} = (z_1, z_2, \dots, z_k)$ independently:

$$p(\mathbf{z} | \mathbf{x}, \bar{\mathbf{z}}; \Theta) = \prod_{i=1}^k p(z_i | \mathbf{x}, \bar{z}_i; \Theta) \quad (1)$$

3.1 Rule-Based Inflection Module

One can predict the inflected form of a target word given its lemma and the source context in two steps: first predict the morphological tag of the target word based on the source context, and then predict the inflected form based on the lemma and morphological tag. The second step can be modeled using traditional inflection models ([Cotterell et al., 2017](#)), while the first step can be performed using rule-based inference based on linguistic knowledge. [McCarthy et al. \(2020\)](#) present a universal morphological (UniMorph) paradigm with universal morphological tags for hundreds of world languages. In UniMorph, the morphological tag of a verb includes information about the tense (past, present, or future), mood (indicative, conditional, imperative, or subjunctive), the number (singular or plural) and person (first, second, or third person) of the subject. The tag of a noun or adjective includes information

about gender (masculine or feminine), number, and grammatical case. Some of these can be inferred from the target lemma (e.g. the gender of a noun) or the source term (e.g. the number of a noun), while some others need to be inferred based on the grammatical function of the source term in the sentence (e.g. grammatical case) or the sentence-level semantics (e.g. mood). Many of the inference rules are shared across a wide range of languages, except for the tense and mood of verbs, as well as the gender and some grammatical cases of nouns and adjectives.

In our rule-based inflection module, we extract the morphological features, part-of-speech tags, and dependency parsing tree of the source sentence using pre-trained Stanza models² and infer the aforementioned classes based on grammar rules and validation examples. The tense and mood of a verb are inferred from the morphological form of the corresponding source term,³ while the number and person of its subject are inferred based on the morphological form of its subject. For nouns and adjectives, the number can be inferred from the morphological form of the source term or modified noun, while the gender can be determined based on the target lemma.

To infer the grammatical case of a noun or adjective, one needs to infer about the grammatical role of the source term in the sentence. For example, in Lithuanian, there are seven main cases, including nominative, genitive, dative, accusative, instrumental, locative, and vocative cases. [Figure 1](#) shows examples of how the case of a Lithuanian noun can be inferred from the dependency parsing tree of the source sentence. Some of the cases can be easily distinguished from the others, while some are more difficult to infer. In this example, the nominative case is comparatively easy to infer – the noun should be in the nominative case when the corresponding source term is the root or subject of the sentence. However, to distinguish between dative, accusative, instrumental, and locative cases, one needs to infer based on the grammatical and semantic role of the source term. In our rule-based module, we only take into account the most common scenarios.⁴

²<https://github.com/stanfordnlp/stanza>

³We ignore tense and mood types that cannot be inferred from the source term.

⁴Our code only includes a few simple inference rules written by non-expert based on the grammar knowledge from Wikipedia pages.

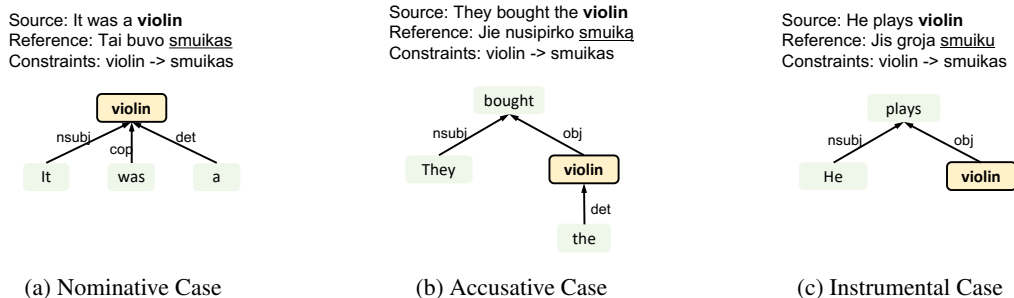


Figure 1: Examples showing how the grammatical case of a target lemma is inferred from the dependency parsing tree of the source sentence. In each example, the reference usage of the target constraint is underlined, and its corresponding source term is boldfaced and highlighted in the yellow, outlined box in each dependency tree. Figure (a) shows an example where the constraint term “smuikas” is used in *nominative* case in the reference, since it is the root in the dependency tree. In Figure (b), the same constraint term is used in *accusative* case in the reference, since it is the *object* of the root verb “bought”. However, not all objects should be used in *accusative* case. As shown in Figure (c), “smuikas” is used in *instrumental* case, since it serves the *instrument* with which the subject performs the action.

Finally, given a lemma and its morphological tag, one can look up its inflected form in a morphological dictionary. We use DEMorphy (Altinok, 2018) for German and Wiktionary⁵ for Lithuanian. Since most Lithuanian nouns follow a set of declension rules,⁶ we inflect Lithuanian nouns based on the rules for lemmas unseen in the dictionary.

3.2 Neural Inflection Module

As prior work shows that BERT-style architectures (Devlin et al., 2019) can encode morphological information in their hidden representations and disambiguate morphologically ambiguous forms via contextualized encoding (Edmiston, 2020), we build the neural-based inflection module as a substitution model and base it on the encoder-decoder Transformer architecture, which embeds the source sentence through the encoder and the target lemmas through the decoder. Next, the decoder predicts the inflected form of each target word in parallel. The inflection module resembles the architecture of the conditional masked language model (CMLM) (Ghazvininejad et al., 2019) but differs in decoder input and output: CMLM takes the target sentence with some tokens masked out as input and is trained to predict only the masked tokens conditioned on unmasked ones, while our inflection module takes target tokens in their lemma forms as input and predicts their inflected forms.

CMLM only allows for one-to-one substitution of subwords. However, in the case of inflection, the number of subwords that constitute a lemma

and its inflected form may differ. To facilitate varying-length substitution, we construct the decoder input by inserting K placeholders at the end of each target lemma. Next, the model predicts the token $t \in \mathcal{V} \cup \{[\text{PLH}]\}$ to be inserted at each input position. If $t = [\text{PLH}]$, we delete the token at this position, otherwise we replace the token at this position with t .⁷

4 Evaluation Test Suites

To evaluate the models’ ability to incorporate diverse types of lemma constraints in different context, we choose the two morphologically complex languages – German and Lithuanian – as the target languages, both of which are fusional languages with strong suffixing. We create two test suites – the English→German health test suite⁸ to evaluate models in the domain adaptation scenario and English→Lithuanian news test suite to test models in the low-resource setting. Different from the automotive test suite of Bergmanis and Pinnis (2021), which contains short sentences (15 tokens per source sentence on average) annotated with limited types of constraints (mostly nouns and proper nouns), our test suites contains longer sentences (20 and 25 tokens per source sentence on average) and diverse types of constraints including adjectives, nouns, proper nouns, and verbs. Different from the upcoming WMT21 terminology task⁹ where

⁷So for instance, given the input “freeze [PLH] [PLH]”, the model could predict the output “fro@@ zen [PLH]”.

⁸To the best of our knowledge, there is no public health (or any non-news) domain MT test set for English→Lithuanian.

⁹<http://statmt.org/wmt21/terminology-task.html>

⁵<https://www.wiktionary.org/>

⁶https://en.wikipedia.org/wiki/Lithuanian_declension

	Source	Constraints	Reference
En-De Health	The routine use of abdominal drainage to reduce postoperative complications after appendectomy for complicated appendicitis is controversial.	abdominal <u>abdominell</u> appendectomy <u>Appendektomie</u> appendicitis <u>Appendizitis</u>	Die routinemäßige Verwendung von <i>abdomineller</i> Drainage zur Verminderung postoperativer Komplikationen nach einer Appendektomie bei komplizierter <u>Appendizitis</u> ist umstritten.
En-Lt News	A fire in 1939 left the building badly damaged, but as Father Johnson ’s parishioners made plans to rebuild , they commissioned the carillon .	Johnson <u>Johnsonas</u> carillon <u>karilionas</u>	1939 m. kilęs gaisras smarkiai apgadino pastatą, tačiau Tėvo <i>Johnsono</i> parapijiečiai planavo jį atstatyti, todėl užsakė <u>karilioną</u> .
	The expert who played the carillon in July called it something else: “A cultural treasure” and “an irreplaceable historical instrument”.	carillon <u>karilionas</u>	Liepos mėnesį <u>karilionu</u> grojęs ekspertas pavadino jį kitaip: “kultūros lobiu” ir “nepakeičiamu istoriniu instrumentu”.

Table 1: Examples from the English→German (En-De) health and English→Lithuanian (En-Lt) news test suites. For En-Lt, we select two examples from the same document. The annotated source terms are boldfaced and the target constraint terms are underlined. Some terms can be copied to the target (e.g. “Lymphödem” and “klinisch” in En-De), while some others need to be inflected in the target sentence (italicized).

	#Sent	#Const	#Const.Inf
Health	3000	4589	802
News	823	374	132

Table 2: Number of sentences (*#Sent*), constraints (*#Const*), and constraints that need to be inflected (*#Const.Inf*) in the health and news test suites.

the terminology translation table includes different forms for a given source term, our test suites only provides terminology translations in lemma forms.

Health Test Suite We construct the health test suite to test the models’ ability to integrate terminology translations for fast domain adaptation. The test set contains English health information text annotated with domain-specific terminology translations and the human-translated sentences in German. We extract English→German test examples from the Himl Test Set,¹⁰ which consists of English health information texts manually translated into German. We extract keyphrases from each source sentence using Yet Another Keyword Extractor (YAKE) (Campos et al., 2020)¹¹ and filter out phrases with high or medium frequency in the training corpora since they are mostly common and domain-generic phrases.¹² We extract

terminology translations from WikiTitles¹³ and an online English-German dictionary,¹⁴ and annotate the keyphrases whose dictionary translations match the reference translation. As shown in Table 1, each source sentence in the test set is annotated with health-related terminology translations in the lemma forms, some of which can be directly copied to the final translation while some need to be inflected based on the context.

News Test Suite The news test suite simulates the scenario where a user looks up keyphrases of a document in a bilingual dictionary and pick the top translation for each keyphrase as a constraint to help low-resource MT. We choose English→Lithuanian as an example of low-resource translation. The test suite is constructed from English→Lithuanian test examples from WMT 2019 news test sets. We first extract keyphrases from each source document using YAKE. Then, we find the top translation of each keyphrase (for many terms there’s only one translation available) in an online dictionary.¹⁵ We filter out the keyphrases whose translations do not match the reference. Table 1 shows two examples from the same document in the test suite. All occurrences of a keyphrase in one document are

¹⁰<http://www.himl.eu/test-sets>

¹¹YAKE extracts n-grams as keyphrases based on word casing, frequency, position, and their sentence context.

¹²We filter out keyphrases with frequency > 100 in the WMT news training data.

¹³<http://data.statmt.org/wikititles/v1/> and <http://data.statmt.org/wikititles/v2/>

¹⁴<https://www.dict.cc/>

¹⁵<https://lithuanian.english-dictionary.help>

annotated with its target translation to encourage consistent translation of keyphrases within a document.¹⁶ Table 2 shows the number of sentences and constraints in each test suite.

5 Experimental Settings

Training Data For English→German (En-De), we use the training corpora from WMT14 (Bojar et al., 2014) and *newstest2013* for validation. For English→Lithuanian, we use the training data from WMT19 (Barrault et al., 2019) and *newsdev2019* as the validation set. For preprocessing, we apply normalization, tokenization, true-casing, and BPE (Sennrich et al., 2016).¹⁷

Baselines We compare our model with the following baselines:

- **Auto-Regressive (AR) baseline** without integrating terminology constraints.
- **AR with Constrained Decoding (CD)** to incorporate hard constraints (Post, 2018).
- **AR with Target Lemma Annotation (TLA)** that integrates lemma constraints as an additional input stream on the source side (Bergmanis and Pinnis, 2021).
- **Non-AutoRegressive (NAR) baseline** based on the EDITOR model (Xu and Carpuat, 2021).
- **NAR with constraints (NAR+C)** that integrates constraints as the initial sequence in EDITOR without explicit inflection.

MT Models All models are based on the *base* Transformer (Vaswani et al., 2017).¹⁸ All models are trained with the Adam optimizer (Kingma and Ba, 2015) with initial learning rate of 0.0005 and effective batch sizes of 32k tokens for AR models and 64k tokens for NAR models for maximum 300k steps.¹⁹ We select the best checkpoint based on validation perplexity. NAR models are trained via sequence-level knowledge distillation (Kim and Rush, 2016). For decoding, we use beam search with a beam size of 4 for AR and AR with TLA, while for AR with CD we use a beam size of 20

¹⁶Interestingly, in Lithuanian, the masculine foreign names are usually translated by appending a suffix to the name to reflect their inflection forms. In this example, the foreign name “Johnson” is translated into “Johnsonas” in the nominative form in the dictionary, while in the reference it becomes “Johnsono” in the genitive form.

¹⁷See preprocessing details and data statistics in Appendix.

¹⁸See more details in Appendix.

¹⁹As shown in prior work (Zhou et al., 2020), the batch sizes for training NAR models are typically larger than the AR model.

as suggested in prior work (Post and Vilar, 2018). To enhance constraint usage in NAR models, we adopt the techniques by Susanto et al. (2020): we prohibit deletions on constraint tokens or insertions within the constraint segments.

Neural Inflection Model Its synthetic training data is derived from the MT parallel data. We first lemmatise and part-of-speech tag the target sentences using Stanza. We then randomly select adjectives, verbs, nouns, and proper nouns from each target sentence and train the inflection module to predict their inflected forms based on their lemma forms and the source sentence. Following Bergmanis and Pinnis (2021), we draw the proportion of words selected in each target sentence randomly from the uniform distribution between (0, 0.4]. For training, we initialize its encoder parameters using the NAR baseline encoder and train it using Adam optimizer with a batch size of 32k tokens for maximum 200k steps.

Evaluation We evaluate translation quality using sacreBLEU (Post, 2018). To evaluate how well the translation preferences are incorporated in the translation outputs, we measure **lemma usage rate** by first lemmatising the translation output and then computing the percentage of lemma terms that appear in the lemmatised output. To evaluate whether the terms are inflected correctly, we measure **term usage accuracy** by matching each lemma constraint with its inflected form in the reference and computing the percentage of reference inflected terms that appear in the translation output.

6 Results and Discussion

Intrinsic Inflection Accuracy To evaluate the quality of the inflection modules, we first compare the inflection accuracy of neural-based and rule-based inflection modules against the term usage accuracy of the TLA model. The rule-based inflection module achieves higher inflection accuracy than the neural-based module on both test suites: the neural-based module obtains 81.2% accuracy on En-De health set and 15.4% accuracy on En-Lt news set, while the rule-based module achieves 87.6% accuracy on En-De and 77.4% accuracy on En-Lt. The rule-based module achieves close accuracy to TLA on En-De (89.2% term usage accuracy) and higher accuracy on En-Lt (67.9% term usage accuracy).

To investigate why the neural-based inflection underperforms the rule-based one, we examine how

	BLEU	Lemma Usage			Term Usage		
		All	No Inf	Inf	All	No Inf	Inf
En-De Health							
AR baseline	31.9	61.2	61.1	61.6	56.7	59.6	43.0
AR w/ CD	33.4	98.6	99.1	96.3	82.6	99.1	4.5
AR w/ TLA	33.8	96.6	97.0	95.0	89.2	94.6	63.6
NAR baseline	31.0	56.1	56.4	54.7	52.8	55.2	41.3
NAR+C	31.1	99.0	99.1	98.5	82.0	99.1	1.4
AR w/ CD + neural	33.3	95.6	95.9	91.1	81.0	91.1	33.3
AR w/ TLA + neural	33.6	94.5	95.1	91.9	85.5	90.2	63.5
NAR+C + neural	30.9	95.6	95.8	94.9	81.1	91.1	33.8
AR w/ CD + rule	33.7	96.8	96.8	97.0	87.3	95.0	51.0
AR w/ TLA + rule	33.9	95.2	95.5	94.1	87.9	92.1	68.0
NAR+C + rule	31.7	97.1	97.0	97.5	87.1	95.0	49.5
En-Lt News							
AR baseline	14.1	64.7	76.9	42.4	55.3	74.0	21.2
AR w/ CD	13.8	89.8	99.6	72.0	65.2	98.8	3.8
AR w/ TLA	14.4	81.5	90.1	65.9	67.9	88.0	31.1
NAR baseline	14.3	59.4	69.0	41.7	52.7	67.8	25.0
NAR+C	14.3	89.8	99.2	72.7	64.7	98.3	3.0
AR w/ CD + neural	13.5	82.4	85.1	77.3	57.2	75.2	24.2
AR w/ TLA + neural	14.2	81.6	86.8	72.0	63.1	78.5	34.8
NAR+C + neural	14.0	83.7	88.0	75.8	58.0	77.7	22.0
AR w/ CD + rule	13.9	93.0	97.5	84.8	75.9	94.2	42.4
AR w/ TLA + rule	14.3	85.3	90.5	75.8	70.3	87.2	39.4
NAR+C + rule	14.3	93.3	97.1	86.4	75.7	94.2	41.7

Table 3: BLEU, lemma, and term usage rates on the En-De health and En-Lt news test suites. For lemma and term usage, we report scores on all constraints (*All*), constraints that require no inflection (*No Inf*), and constraints that require inflection (*Inf*). We boldface the highest scores and their ties based on the paired bootstrap significance test (Clark et al., 2011) with $p < 0.05$.

the training and validation perplexity changes over the number of training epochs (see Appendix). On both languages, the validation perplexity stops decreasing after a few training epochs (10 epochs for En-De and 20 epochs for En-Lt) while the training perplexity decreases very slowly. The final training perplexity remains at around 5.1 on En-De and 5.7 on En-Lt, which is high considering the number of possible inflection forms given a German or Lithuanian lemma. This indicates that the neural-based module does not learn generalizable inflection rules from the data effectively.

End-to-End MT Evaluation Table 3 shows the impact of rule-based and neural-based inflection modules on top of a range of AR and NAR baselines. NAR baselines without constraints achieves competitive BLEU to the AR baseline on En-Lt and slightly lower BLEU on En-De, as in Xu and Carpuat (2021). Given lemma constraints, AR with

CD without inflection obtains lower term usage accuracy and lower BLEU than AR with TLA, as in Bergmanis and Pinnis (2021). Similar to AR with CD, NAR+C without inflection obtains lower term usage and close or lower BLEU than AR with TLA.

Adding rule-based inflection helps all models leverage lemma constraints more accurately. On En-De, it significantly improves term usage accuracy of AR with CD by +4.7% and NAR+C models by +5.1%.²⁰ On En-Lt, it significantly improves both the lemma usage rate and term usage accuracy of AR with CD (+3.2% on lemma usage and +10.7% on term usage) and NAR+C (+3.5% on lemma usage and +11.0% on term usage). Remarkably, it also improves the term accuracy of En-Lt AR with TLA, which is already trained to inflect the target lemma constraints. When evaluating only on constraints

²⁰All mentions of significance are based on the paired bootstrap test (Clark et al., 2011) with $p < 0.05$.

that require inflection, the rule-based modules improves by 4.4–8.3% on TLA, 38.6–46.5% on CD, and 38.7–48.1% on NAR+C. As expected based on inflection accuracy results, rule-based modules outperform neural-based ones across the board. These improvements in term usage preserve or slightly improve BLEU.²¹, as can be expected since the constraints only constitute a small portion of the tokens in the translation outputs. Overall, these results indicate that our proposed framework is model-agnostic and supports our hypothesis that the lemma constraints can be effectively inflected based on the source context alone.

We now compare our framework against TLA. Rule-based inflection combined with NAR+C achieves close lemma and term usage rates ($\Delta \leq 2\%$) to TLA on En-De, +11.8% higher lemma usage, and +7.8% higher term usage accuracy on En-Lt (the improvements are significant). On En-Lt, the largest improvements are on constraints that require inflection: +20.5% on lemma usage and +10.6% on term usage. Incorporating the constraints preserves translation quality, with no significant difference in BLEU. Overall, these results show the benefits of integrating linguistic knowledge via rule-based inflection over purely data-driven approaches. Our approach is also more adaptive, as NAR+C with rule-based inflection does not require re-training the whole NMT model to incorporate new lemma terms. Instead, new terms can be incorporated by updating the morphological dictionary used in the inflection module.

Cost Trade-offs Implementing the rule-based inflection module for the first target language (Lithuanian) took around 6 hours (including the time for learning the grammar knowledge from Wikipedia) by a computer scientist without prior knowledge of the target language nor formal linguistics training. The second language (German) implementation took only 3 hours, since some rules are shared across languages. By contrast, the neural-based module was implemented in about 3 hours but took around 38 hours to train a single model for one language pair on 2 GeForce GTX 1080 Ti GPUs. While these numbers do not provide a controlled comparison, they highlight that the rule-based module is relatively simple to build, as it can be done for both languages in 7-15% of the time required to train the neural model.

²¹The improvements on BLEU is statistically significant for NAR+C on En-De, but not for other models.

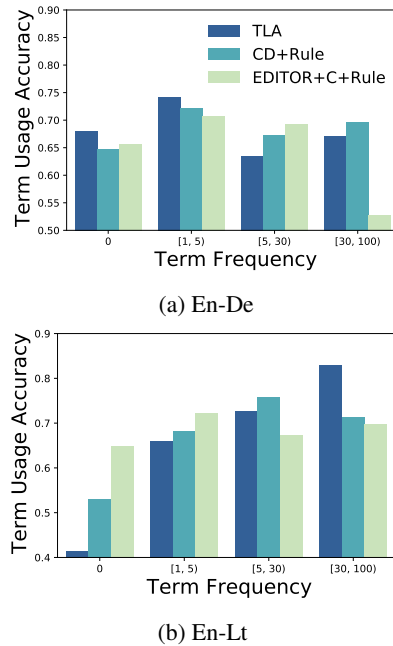


Figure 2: Term usage accuracy of TLA, CD + rule, and NAR+C + rule binned by training set frequency.

Term Frequency We analyze where rule-based inflection helps the most by computing the term usage accuracy on terms in different frequency bucket. As shown in Figure 2, the trends are different on En-De and En-Lt. On En-De, CD + rule slightly improves TLA on terms with frequency between [5, 100) instead of the rare terms. One reason is that the German morphological dictionary that we use to determine the gender of a word and its inflection forms only covers around 70% of the constraint terms in the health test suite. In addition, NAR+C + rule underperforms CD + rule on some constraint terms with frequency between [30, 100). This might be a side effect of knowledge distillation, which yields frequent errors for words that are rare in the training data (Ding et al., 2021). In En-Lt test set, 68% of the constraint terms are used in the inflection forms that are unseen in the training data. As shown in the figure, both CD + rule and NAR+C + rule bring substantial improvements over TLA on terms that are unseen in the training data. This is because most Lithuanian nouns and adjectives are inflected based on a fixed set of rules, thus even when the target lemma is unseen in the training data or morphological dictionary, it can still be inflected correctly. As a result, the rule-based inflection module can effectively incorporate linguistic knowledge in translation models and thus generalizes better to rare and unseen terms.

source	Jim Furyk’s side need eight points from Sunday’s 12 singles matches to retain the trophy .
reference	Jimo Furyko komandai reikia gauti aštuonis taškus sekmadienio 12 vienetų mačiuose, kad išsaugotų <u>trofėjų</u> .
constraints	trophy : <u>trofėjus</u>
reference inflection	trophy : <u>trofėjų</u> (singular, accusative)
TLA	Jim Furyk ’s pusėje reikia aštuonių taškų iš sekmadienio 12 pažintys rungtynes išlaikyti <u>trofėjus</u> .
TLA + rule	Jim Furyk ’s pusėje reikia aštuonių taškų iš sekmadienio 12 pažintys rungtynes išlaikyti <u>trofėjų</u> .
source	In December 2017, he was accused of assaulting his father, Todd Palin .
reference	2017 m. gruodžio mėnesį jis buvo apkaltintas smurtu prieš savo tėvą Toddą <u>Paliną</u> .
constraints	Palin : <u>Palinas</u>
reference inflection	Palin : <u>Paliną</u> (singular, accusative)
NAR+C	2017 m. gruodžio mėn. jis buvo apkaltintas užpuolimu jo tėvas Toddas <u>Palinas</u> .
NAR+C + rule	2017 m. gruodžio mėn. jis buvo apkaltintas užpuolęs tėvą Toddą <u>Paliną</u> .

Table 4: Translation examples comparing TLA + rule against TLA, and NAR+C + rule against NAR+C on En-Lt. We boldface the source terms with translation constraints and underline the target constraint terms used in the reference and translation outputs.

Qualitative Analysis We examine a few randomly selected translation examples from TLA, NAR+C, and their counterparts with rule-based inflection. As shown in Table 4, TLA tends to copy constraint terms that are infrequent in the training data, and adding the rule-based inflection module helps TLA inflect the term correctly instead. In NAR+C models, the inflection module also improves the translation of the context around constraint terms, while the vanilla NAR+C model is prone to compounding errors caused by the uninflected constraints.

7 Conclusion

We introduced a modular framework for leveraging terminology constraints provided in lemma forms in neural machine translation. The framework is based on a novel cross-lingual inflection module that inflects the target lemma constraints given source context and an NMT model that integrates the inflected constraints in the output. We showed that our framework can be flexibly applied to different types of inflection modules, including rule-based and neural-based ones, and different NMT models, including autoregressive and non-autoregressive ones, with minimal training costs. Results on the English-German health and English-Lithuanian test suites showed that the linguistically motivated rule-based inflection

module helps NMT models incorporate terminology constraints more accurately than both neural-based inflection and the existing end-to-end approach to incorporating lemma constraints. This work opens future avenues for further improving the inflection module by combining linguistic knowledge with data-driven approaches. Future work is needed to explore the strengths and weaknesses of this framework for languages with a broader range of morphological properties.

Acknowledgement

We thank Sweta Agrawal, Hal Daumé III, the anonymous reviewers, the CLIP lab at UMD for their helpful and constructive comments, and Yuxin Xiong for her help on error analysis of German outputs. This research is supported in part by the Amazon Web Services Machine Learning Research Award and by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract #FA8650-17-C-9117. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

- Duygu Altinok. 2018. [Demorphy, german language morphological analyzer](#). *CoRR*, abs/1803.00902.
- Duygu Ataman and Marcello Federico. 2018. [Compositional representation of morphologically-rich input for neural machine translation](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 305–311, Melbourne, Australia. Association for Computational Linguistics.
- Tamali Banerjee and Pushpak Bhattacharyya. 2018. [Meaningless yet meaningful: Morphology grounded subword-level NMT](#). In *Proceedings of the Second Workshop on Subword/Character Level Models*, pages 55–60, New Orleans. Association for Computational Linguistics.
- Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. [Findings of the 2019 conference on machine translation \(WMT19\)](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.
- Toms Bergmanis and Mārcis Pinnis. 2021. [Facilitating terminology translation with target lemma annotations](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3105–3111, Online. Association for Computational Linguistics.
- Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. [Findings of the 2014 workshop on statistical machine translation](#). In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.
- Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alipio Jorge, Célia Nunes, and Adam Jatowt. 2020. [Yake! keyword extraction from single documents using multiple local features](#). *Information Sciences*, 509:257–289.
- Fabienne Cap, Alexander Fraser, Marion Weller, and Aoife Cahill. 2014. [How to produce unseen teddy bears: Improved morphological processing of compounds in SMT](#). In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 579–587, Gothenburg, Sweden. Association for Computational Linguistics.
- Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. [Better hypothesis testing for statistical machine translation: Controlling for optimizer instability](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 176–181, Portland, Oregon, USA. Association for Computational Linguistics.
- Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. [The CoNLL-SIGMORPHON 2018 shared task: Universal morphological reinflection](#). In *Proceedings of the CoNLL-SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels. Association for Computational Linguistics.
- Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. [CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages](#). In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30, Vancouver. Association for Computational Linguistics.
- 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.
- Liang Ding, Longyue Wang, Xuebo Liu, Derek F. Wong, Dacheng Tao, and Zhaopeng Tu. 2021. [Understanding and improving lexical choice in non-autoregressive translation](#). In *International Conference on Learning Representations*.
- 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.
- Daniel Edmiston. 2020. [A systematic analysis of morphological content in BERT models for multiple languages](#). *CoRR*, abs/2004.03032.
- Manaal Faruqui, Yulia Tsvetkov, Graham Neubig, and Chris Dyer. 2016. [Morphological inflection generation using character sequence to sequence learning](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 634–643, San Diego, California. Association for Computational Linguistics.

- Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. [Mask-predict: Parallel decoding of conditional masked language models](#). 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 6112–6121, Hong Kong, China. Association for Computational Linguistics.
- Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. 2018. [Non-autoregressive neural machine translation](#). In *International Conference on Learning Representations*.
- Jiatao Gu, Changhan Wang, and Junbo Zhao. 2019. [Levenshtein transformer](#). In *Advances in Neural Information Processing Systems 32*, pages 11181–11191. Curran Associates, Inc.
- 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.
- Matthias Huck, Simon Riess, and Alexander Fraser. 2017. [Target-side word segmentation strategies for neural machine translation](#). In *Proceedings of the Second Conference on Machine Translation*, pages 56–67, Copenhagen, Denmark. Association for Computational Linguistics.
- Ronald M. Kaplan and Martin Kay. 1994. [Regular models of phonological rule systems](#). *Computational Linguistics*, 20(3):331–378.
- Yova Kementchedjheva, Johannes Bjerva, and Isabelle Augenstein. 2018. [Copenhagen at CoNLL–SIGMORPHON 2018: Multilingual inflection in context with explicit morphosyntactic decoding](#). In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 93–98, Brussels. Association for Computational Linguistics.
- Yoon Kim and Alexander M. Rush. 2016. [Sequence-level knowledge distillation](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.
- Diederik P. Kingma and Jimmy Ba. 2015. [Adam: A method for stochastic optimization](#). In *Proceedings of the 3th International Conference on Learning Representations*, San Diego, CA, USA.
- Philipp Koehn and Kevin Knight. 2003. [Empirical methods for compound splitting](#). In *10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary. Association for Computational Linguistics.
- Kimmo Koskenniemi. 1984. [A general computational model for word-form recognition and production](#). In *Proceedings of the 4th Nordic Conference of Computational Linguistics (NODALIDA 1983)*, pages 145–154, Uppsala, Sweden. Centrum för datorlingvistik, Uppsala University, Sweden.
- Jason Lee, Elman Mansimov, and Kyunghyun Cho. 2018. [Deterministic non-autoregressive neural sequence modeling by iterative refinement](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, Brussels, Belgium. Association for Computational Linguistics.
- Xuezhe Ma, Chunting Zhou, Xian Li, Graham Neubig, and Eduard Hovy. 2019. [FlowSeq: Non-autoregressive conditional sequence generation with generative flow](#). 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 4282–4292, Hong Kong, China. Association for Computational Linguistics.
- Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Natalya Krizhanovskiy, Andrew Krizhanovskiy, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. [UniMorph 3.0: Universal Morphology](#). In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.
- Toan Q. Nguyen and David Chiang. 2018. [Improving lexical choice in neural machine translation](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 334–343. Association for Computational Linguistics.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- 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.
- Ofir Press and Lior Wolf. 2017. [Using the output embedding to improve language models](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Compu-*

- tational*, pages 157–163. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Neural machine translation of rare words with subword units](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725. Association for Computational Linguistics.
- Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019. [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.
- 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.
- Aleš Tamchyna, Marion Weller-Di Marco, and Alexander Fraser. 2017. [Modeling target-side inflection in neural machine translation](#). In *Proceedings of the Second Conference on Machine Translation*, pages 32–42, Copenhagen, Denmark. Association for Computational Linguistics.
- Ferhan Ture, Douglas W. Oard, and Philip Resnik. 2012. [Encouraging consistent translation choices](#). In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 417–426, Montréal, Canada. Association for Computational Linguistics.
- Aaron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George van den Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, Norman Casagrande, Dominik Grewe, Seb Noury, Sander Dieleman, Erich Elsen, Nal Kalchbrenner, Heiga Zen, Alex Graves, Helen King, Tom Walters, Dan Belov, and Demis Hassabis. 2018. [Parallel WaveNet: Fast high-fidelity speech synthesis](#). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3918–3926, Stockholm, Sweden. PMLR.
- 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, pages 5998–6008, Long Beach, CA, USA. Curran Associates, Inc.
- David Wan, Chris Kedzie, Faisal Ladhak, Marine Carpuat, and Kathleen McKeown. 2020. [Incorporating terminology constraints in automatic post-editing](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1193–1204, Online. Association for Computational Linguistics.
- Weijia Xu and Marine Carpuat. 2021. [EDITOR: An Edit-Based Transformer with Repositioning for Neural Machine Translation with Soft Lexical Constraints](#). *Transactions of the Association for Computational Linguistics*, 9:311–328.
- Chunting Zhou, Jiatao Gu, and Graham Neubig. 2020. [Understanding knowledge distillation in non-autoregressive machine translation](#). In *International Conference on Learning Representations*.

A Data Preprocessing

For preprocessing, we apply normalization, tokenization, true-casing, and BPE (Sennrich et al., 2016) with 37,000 and 24,500 merging operations for En-De and En-Lt. Table 5 shows the provenance and statistics of the preprocessed data.

B Model and Training Details

All models are based on the *base* Transformer (Vaswani et al., 2017) with $d_{\text{model}} = 512$, $d_{\text{hidden}} = 2048$, $n_{\text{heads}} = 8$, $n_{\text{layers}} = 6$, and $p_{\text{dropout}} = 0.3$. We tie the source and target embeddings with the output layer weights (Press and Wolf, 2017; Nguyen and Chiang, 2018). We add dropout to embeddings (0.1) and label smoothing (0.1). All models are trained with the Adam optimizer (Kingma and Ba, 2015) with initial learning rate of 0.0005 and effective batch sizes of $32k$ tokens for AR models and $64k$ tokens for NAR models for maximum 300,000 steps.²² We select the best checkpoint based on validation perplexity. Following Xu and Carpuat (2021), we train NAR models using sequence-level knowledge distillation: we replace the reference sentences in the training data with translation outputs from the AR models. To train the neural-based inflection module, we initialize its encoder parameters using the NAR baseline encoder and train it using Adam optimizer with a batch size of $32k$ tokens for maximum 200,000 steps. Models are trained on 2 GeForce GTX 1080 Ti GPUs. Table 6 shows the number of parameters in each model.

	Train	Valid	Provenance
En-De	3,961k	3,000	WMT14
En-Lt	1,612k	1,964	WMT19

Table 5: Number of sentence pairs and provenance of the training and validation data.

C Evaluation Metric

We evaluate translation quality using sacre-BLEU (Post, 2018).²³

²²As shown in prior work, the batch sizes for training non-autoregressive models are typically larger than the AR model (Zhou et al., 2020).

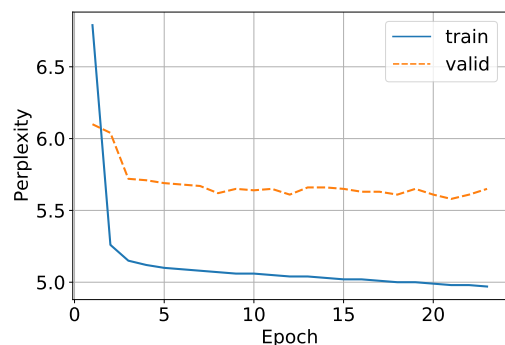
²³BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.2.11

	Model Size (M)
En-De	
AR	65
AR w/ CD	65
AR w/ TLA	65
NAR	91
rule-based inflection	0
neural-based inflection	86
En-Lt	
AR	57
AR w/ CD	57
AR w/ TLA	58
NAR	84
rule-based inflection	0
neural-based inflection	72

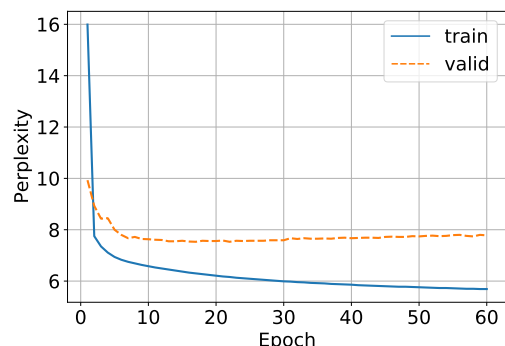
Table 6: Model sizes (M) for the AR, NAR, and inflection models.

D Learning Curves

Figure 3 shows the learning curves of En-De and En-Lt neural-based inflection modules in terms of training and validation perplexity.



(a) En-De



(b) En-Lt

Figure 3: Training and validation perplexity of the En-De and En-Lt neural-based inflection modules.