Facilitating Terminology Translation with Target Lemma Annotations

Toms Bergmanis[†‡] and **Mārcis Pinnis**[†‡]

[†]Tilde / Vienības gatve 75A, Riga, Latvia
[‡]Faculty of Computing, University of Latvia / Raiņa bulv. 19, Riga, Latvia
{firstname.lastname}@tilde.lv

Abstract

Most of the recent work on terminology integration in machine translation has assumed that terminology translations are given already inflected in forms that are suitable for the target language sentence. In day-to-day work of professional translators, however, it is seldom the case as translators work with bilingual glossaries where terms are given in their dictionary forms; finding the right target language form is part of the translation process. We argue that the requirement for apriori specified target language forms is unrealistic and impedes the practical applicability of previous work. In this work, we propose to train machine translation systems using a source-side data augmentation method[1] that annotates randomly selected source language words with their target language lemmas. We show that systems trained on such augmented data are readily usable for terminology integration in real-life translation scenarios. Our experiments on terminology translation into the morphologically complex Baltic and Uralic languages show an improvement of up to 7 BLEU points over baseline systems with no means for terminology integration and an average improvement of 4 BLEU points over the previous work. Results of the human evaluation indicate a 47.7% absolute improvement over the previous work in term translation accuracy when translating into Latvian.

1 Introduction

Translation into morphologically complex languages involves 1) making a lexical choice for a word in the target language and 2) finding its morphological form that is suitable for the morpho-syntactic context of the target sentence. Most of the recent work on terminology translation, however, has assumed that the correct morphological forms are apriori known (Hokamp and Liu, 2017; Post and Vilar, 2018; Hasler et al., 2018; Dinu et al., 2019; Song et al., 2020; Susanto et al., 2020; Dougal and Lonsdale, 2020). Thus previous work has approached terminology translation predominantly as a problem of making sure that the decoder's output contains *lexically* and *morphologically* prespecified target language terms. While useful in some cases and some languages, such approaches come short of addressing terminology translation into morphologically complex languages where each word can have many morphological surface forms.

For terminology translation to be viable for translation into morphologically complex languages, terminology constraints have to be *soft*. That is, terminology translation has to account for various natural language phenomena, which cause words to have more than one manifestation of their root morphemes. Multiple root morphemes complicate the application of *hard* constraint methods, such as constrained-decoding (Hokamp and Liu, 2017). That is because even after the terminology constraint is striped from the morphemes that encode all grammatical information, the remaining root morphemes still can be too restrictive to be used as hard constraints because, for many words, there can be more than one root morpheme possible. An illustrative example is the consonant mutation in the Latvian noun *vācietis* ("the German") which undergoes the mutation t→š, thus yielding two variants of its root morpheme *vācieš-* and *vāciet-* (Bergmanis, 2020). If either of the forms is used as a hard constraint for constrained decoding, the other one is excluded from appearing in the sentence's translation.

We propose a necessary modification for the method introduced by Dinu et al. (2019), which allows training neural machine translation (NMT)

[1]Relevant materials and code: https://github.com/tilde-nlp/terminology_translation

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 3105–3111
April 19 - 23, 2021.

EN Src.:	faulty engine or in transmission[..]
LV Trg.:	atteice dzinējā vai transmisijas [..]
ETA:	faulty|w engine|s dzinēj**ā**|t or|w transmission|s transmisij**as**|t [..]
TLA:	faulty|w engine|s dzinēj**s**|t or|w transmission|s transmisij**a**|t [..]

Table 1: Examples of differences in input data in ETA (Dinu et al., 2019) and TLA (this work). Differences of inline annotations are marked in bold. |w, |s, |t denote the values of the additional input stream and stand for regular words, source language annotated words, target language annotations respectively.

	Train	Test	
		ATS	WMT17+IATE
EN-DE	27.6M	768	581
EN-ET	2.4M	768	-
EN-LV	22.6M	768	-
EN-LT	22.1M	768	-

Table 2: Training and evaluation data sizes in numbers of sentences. WMT2017 + IATE stands for the English-German test set from the news translation task of WMT2017 which is annotated with terminology from the IATE terminology database.

systems that are capable of applying terminology constraints: instead of annotating source-side terminology with their exact target language translations, we annotate randomly selected source language words with their target language lemmas. First of all, preparing training data in such a way relaxes the requirement for access to bilingual terminology resources at the training time. Second, we show that the model trained on such data does not learn to simply *copy* inline annotations as in the case of Dinu et al. (2019), but learns *copy-and-inflect* behaviour instead, thus addressing the need for *soft* terminology constraints.

Our results show that the proposed approach not only relaxes the requirement for apriori specified target language forms but also yields substantial improvements over the previous work (Dinu et al., 2019) when tested on the morphologically complex Baltic and Uralic languages.

2 Method: Target Lemma Annotations

To train NMT systems that allow applying terminology constraints Dinu et al. (2019) prepare training data by amending source language terms with their exact target annotations (**ETA**). To inform the NMT model about the nature of each token (i.e., whether it is a source language term, its target language translation or a regular source language word), the authors use an additional input stream—source-side factors (Sennrich and Haddow, 2016). Their method, however, is limited to cases in which the provided annotation matches the required target form and can be copied verbatim, thus performing poorly in cases where the surface forms of terms in the target language differ from those used to annotate source language sentences (Dinu et al., 2019). This constitutes a problem for the method's practical applicability in real-life scenarios. In this work, we propose two changes to the approach of Dinu et al. (2019). **First**, when preparing training data, instead of using terms found in either IATE[2] or Wiktionary as done by Dinu et al. (2019), we annotate random source language words. This relaxes the requirement for curated bilingual dictionaries for training data preparation. **Second**, rather than providing exactly those target language forms that are used in the target sentence, we use target lemma annotations (**TLA**) instead (see Table 1 for examples). We hypothesise that in order to benefit from such annotations, the NMT model will have to learn *copy-and-inflect* behaviour instead of simple *copying* as proposed by Dinu et al. (2019).

Our work is similar to work by Exel et al. (2020) in which authors also aim to achieve *copy-and-inflect* behaviour. However, authors limit their annotations to only those terms for which their base forms differ by no more than two characters from the forms required in the target language sentence. Thus wordforms undergoing longer affix change or inflections accompanied by such linguistic phenomena as consonant mutation, consonant gradation or other stem change are never included in training data.

3 Experimental Setup

Languages and Data. As our focus is on morphologically complex languages, in our experiments we translate from English into Latvian and Lithuanian (Baltic branch of the Indo-European language family) as well as Estonian (Finnic branch of the Uralic language family). For comparability with the previous work, we also use English-German (Germanic branch of the Indo-European language family). For all language pairs, we use all data that is available in the Tilde Data Libarary with an exception for English-Estonian for which

[2] https://iate.europa.eu

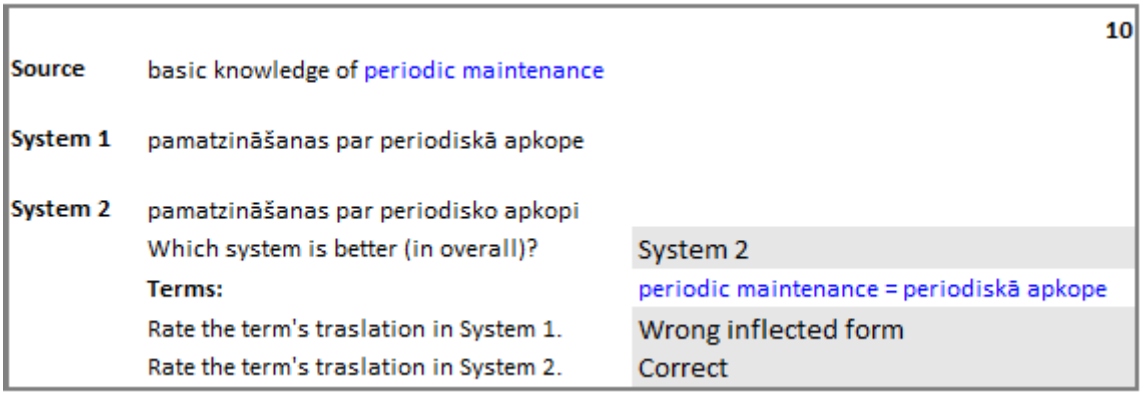

Figure 1: Example of forms used in human evaluation.

we use data from WMT 2018. The size of the parallel corpora after pre-processing using the Tilde MT platform (Pinnis et al., 2018) and filtering tools (Pinnis, 2018) is given in Table 2.

To prepare data with TLA, we first lemmatise and part-of-speech (POS) tag the target language side of parallel corpora. For lemmatisation and POS tagging, we use pre-trained Stanza[3] (Qi et al., 2020) models. We then use *fast_align*[4] (Dyer et al., 2013) to learn word alignments between the target language lemmas and source language inflected words. We only annotate verbs or nouns. To generate sentences with varying proportions of annotated and unannotated words, we first generate a sentence level annotation threshold uniformly at random from the interval $[0.6, 1.0)$. Similarly, for each word in the source language sentence, we generate another number uniformly at random from the interval $[0.0, 1.0)$. If the latter is larger than the sentence level annotation threshold, we annotate the respective word with its target language lemma. We use the original training data and annotated data with a proportion of 1:1. We follow Dinu et al. (2019) to prepare ETA and replicate their results.

For validation during training, we use development sets from the WMT news translation shared tasks. For EN-ET and EN-DE, we used the data from WMT 2018, for EN-LV – WMT 2017, and for EN-LT – WMT 2019.

MT Model and Training. For the most part, we use the default configuration of the Transformer (Vaswani et al., 2017) NMT model implementation of the Sockeye NMT toolkit (Hieber et al.). The exception is the use of source-side factors (Sennrich and Haddow, 2016) with the dimensionality of 8 for systems using inline target lemma annotations. We train all models using early stopping with the patience of 10 based on their development set perplexity (Prechelt, 1998).

Evaluation Methods and Data. In previous work, methods were tested on general domain data[5] annotated with exact surface forms of general-domain words from IATE and Wiktionary. Although data constructed in such a way is not only artificial but also gives an oversimplified view on terminology translation, we do use the data from IATE to validate our re-implementation of the method from Dinu et al. (2019). Other than that, we test on the Automotive Test Suite[6] (ATS): a data set containing translations of the same 768 sentences in English, Estonian, German, Latvian, and Lithuanian. ATS contains about 1.1k term occurrences from a glossary prepared by professional translators. When annotating terms in the source text, we use only the dictionary forms of term translations, since in practical applications having access to the correct inflections (surface forms) is unrealistic.

We compare our work with an NMT system without means for terminology integration (**Baseline**) and the previous work by Dinu et al. (2019) (**ETA**). Although our preliminary experiments with constrained decoding (Post and Vilar, 2018) (**CD**) confirmed the findings by Dinu et al. (2019) that strict enforcement of constraints leads to lower-than-baseline quality, we nevertheless include them for completeness sake.

Similarly to the previous work, we use two auto-

[3] https://github.com/stanfordnlp/stanza
[4] https://github.com/clab/fast_align
[5] https://github.com/mtresearcher/terminology_dataset
[6] https://github.com/tilde-nlp/terminology_translation

	IATE EN-DE		Automotive Test Suite EN-DE		EN-ET		EN-LV		EN-LT	
	BLEU	Acc.	BLEU	Acc.	BLEU	Acc.	BLEU	Acc.	BLEU	Acc.
Baseline	29.7	81.7	26.5	46.2	19.6	46.7	30.6	62.2	25.3	51.2
CD	28.5	**99.7**	22.9	**99.7**	14.9	**98.0**	23.5	**99.3**	18.1	**98.9**
ETA	**29.9**	96.2	33.2†	94.0	17.8	92.4	27.4	93.4	28.8†	89.7
TLA	29.5	96.5	**33.5**†	94.0	**21.0**†‡	87.2	**35.0**†‡	92.0	**30.1**†‡	90.3

Table 3: Results of automatic evaluation metrics BLEU and term translation accuracy (Acc.). The numerically highest score in each column is given in bold; † and ‡ indicate statistically significant improvements of BLEU over Baseline and ETA respectively (all $p < 0.05$) .

	Correct	Wrong lexeme	Wrong inflect.	Other	κ_{free}
Basel.	55.1	42.9	1.4	0.7	0.95
ETA	45.2	7.9	44.9	2.0	0.87
TLA	92.9	5.1	1.4	0.7	0.98

Baseline	**Equal**	**TLA**	κ_{free}
3.0	58.0	39.0	0.65
ETA	**Equal**	**TLA**	κ_{free}
3.0	36.0	61.0	0.81

Table 4: Results of human evaluation: term (on the left) and sentence (on the right) translation quality judgements in %. Sentence comparison is pairwise contrasting TLA vs Baseline and TLA vs ETA. κ-free: inter-annotator agreement according to free marginal kappa (Randolph, 2005).

matic means for evaluation: BLEU (Papineni et al., 2002) and lemmatised term exact match accuracy. We use BLEU as an extrinsic evaluation metric as we expect that, when successful, the methods for terminology translation should yield substantial overall translation quality improvements due to correctly translated domain-specific terms. For significance testing, we use pairwise bootstrap resampling (Koehn, 2004). We use lemmatised term exact match accuracy as an intrinsic metric because it directly measures the adequacy of terminology translation (i.e., whether or not the correct lexeme appears in the target sentence).

We are aware that the automatic evaluation methods are merely an approximation of translation quality. For example, we use lemmatised term exact match accuracy to measure term use in target language translations; however, it does not capture whether the term is inflected correctly. Thus human evaluation is in place. We use the EN-LV language pair to compare TLA against baseline and ETA. We use a 100 sentences large randomly selected ATS subset that contains 147 terms of the original test suite. We employ four professional translators and Latvian native speakers to compare each system's translations according to their overall translation quality and judge individual term translation quality. Specifically, given the original sentence and its two translations (in a randomised order), raters are asked to answer "which system's translation is better overall?". Raters are also given a list of the terms being evaluated and their reference translations (from the term collection) and are asked to classify translations as either "Correct", "Wrong lexeme", "Wrong inflection", or "Other". Figure 1 gives an example of the forms presented to raters during the human evaluation of term and overall translation quality. We report inter-annotator agreement using free marginal kappa, κ_{free} (Randolph, 2005).

4 Results

Automatic Evaluation. We first validate our re-implementation of ETA by testing on the English-German WMT 2017 test set annotated with terms from IATE as used by Dinu et al. (2019). Results (see columns 2 and 3 of Table 3) are similar to those of the previous work: on this data set, ETA yields minor translation quality improvements over the baseline (+0.2 BLEU) and considerable improvement (+14.5%) in term translation accuracy.

When evaluated on the ATS, systems using TLA always yield results that are better than the baseline both in terms of BLEU scores (+1.4–7 BLEU) and term translation accuracy (29.8%–47.8%) (see columns 4-11 of Table 3). Results also show that when compared to ETA, systems integrating terminology using TLA achieve statistically significant improvements in terms of BLEU scores for three out of four languages-pairs. An exception is EN-DE, for which both systems, ETA and TLA, perform similarly. Analysing reference translations of the EN-DE language pair, we find that as many

as 87% of the German terms are used in their dictionary forms, which explains the comparable performance of systems trained using ETA and TLA on EN-DE.

Results also confirm the finding of the previous work by Dinu et al. (2019) and Exel et al. (2020), that the strict enforcement of constraints by constrained decoding leads to lower-than-baseline BLEU scores on all data sets for all languages. BLEU scores are abysmal when translating into the morphologically complex languages as for these languages citation form seldom happens to be the form required in the target language sentence. This result further illustrates why terminology constraints have to be *soft* when translating into morphologically complex languages.

Human Evaluation. Results of human evaluation of EN-LV systems are summarised in Table 4. First, we note that on this dataset, the baseline system translates terms correctly 55% of the time, yet it makes mistakes by choosing the wrong lexeme for most of the other cases (Table 4, left). The system using ETA, on the other hand, has a much lower rate of correctly translated terms – 45%, which roughly corresponds to the proportion of Latvian terms in the reference translations that are used in their dictionary forms (47%). The remaining cases are mistranslated by choosing the wrong inflected form. The system using TLA, in comparison, does very well as it gets terminology translations right 93% of the time. Examining the cases where terms had been mistranslated by choosing the wrong lexeme, we find that most of these cases are multi-word terms with some other word inserted between their constituent parts. The high κ-free values indicate almost perfect inter-annotator agreement suggesting that the task of term translation quality evaluation has been easy and results are reliable.

The overall sentence translation quality judgements (Table 4, right) also favour translations produced by the system using TLA deeming it better than or on par with the baseline system and system using ETA 97% of the time. The system using TLA is strictly favoured over its ETA counterpart for 61% of the translations. Again, annotators have reached an almost perfect agreement ($\kappa_{\text{free}} = 0.81$) when comparing the systems using TLA and ETA, suggesting that the task has been easy. These results clearly show that at least for the EN-LV language pair and the test set considered here, systems using TLA improve term translation quality by correctly choosing adequate translations and morpho-syntactically appropriate inflections.

Productivity of NMT models. Terminology translation frequently involves the translation of niche lexemes with rare or even unseen inflections. Thus the model's ability to generate novel wordforms is critical for high-quality translations. To verify if our NMT models are lexically and morphologically productive, we analysed Latvian translations of ATS produced by the system using TLA and looked for wordforms that are not present in either source or target language side of the training data. We found 72 such wordforms. Of those 45 or 62.5% were valid wordforms that were not present in training data, of which 28 were novel inflections related to ATS terminology use, while the remaining 17 where novel forms of general words. We interpret this as *some* evidence that the NMT model, when needed, generates novel wordforms. The remaining 27 or 37.5% were not valid, albeit sometimes plausible, Latvian language words, common types of errors being literal translations and transliterations of English words as well as words that would have been correct, if not for errors with consonant mutation.

5 Conclusions

We proposed TLA—a flexible and easy-to-implement method for terminology integration in NMT. Using TLA does not require access to bilingual terminology resources at system training time as it annotates ordinary words with lemmas of their target language translations. This simplifies data preparation greatly and also relaxes the requirement for apriori specified target language forms during the translation, making our method practically viable for terminology translation in real-life scenarios. Results from experiments on three morphologically complex languages demonstrated substantial and systematic improvements over the baseline NMT systems without means for terminology integration and the previous work both in terms of automatic and human evaluation judging term and overall translation quality.

Acknowledgements

This research has been supported by the ICT Competence Centre (www.itkc.lv) within the project "2.2. Adaptive Multimodal Neural Machine Translation" of EU Structural funds, ID no 1.2.1.1/18/A/003.

References

Toms Bergmanis. 2020. *Methods for morphology learning in low(er)-resource scenarios*. Ph.D. thesis, The University of Edinburgh.

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.

Duane K. Dougal and Deryle Lonsdale. 2020. Improving NMT quality using terminology injection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4820–4827, Marseille, France. European Language Resources Association.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Miriam Exel, Bianka Buschbeck, Lauritz Brandt, and Simona Doneva. 2020. Terminology-constrained neural machine translation at SAP. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 271–280, Lisboa, Portugal. European Association for Machine Translation.

Eva Hasler, Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne. 2018. Neural machine translation decoding with terminology constraints. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 506–512, New Orleans, Louisiana. Association for Computational Linguistics.

Felix Hieber, Tobias Domhan, Michael Denkowski, and David Vilar. Sockeye 2: A toolkit for neural machine translation. In *22nd Annual Conference of the European Association for Machine Translation*, page 457.

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.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 388–395.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Mārcis Pinnis. 2018. Tilde's parallel corpus filtering methods for wmt 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 939–945.

Mārcis Pinnis, Andrejs Vasiļjevs, Rihards Kalniņš, Roberts Rozis, Raivis Skadiņš, and Valters Šics. 2018. Tilde MT platform for developing client specific MT solutions. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

Lutz Prechelt. 1998. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.

Justus J Randolph. 2005. Free-marginal multirater kappa (multirater κfree): An alternative to fleiss' fixed-marginal multirater kappa. In *Presented at the Joensuu Learning and Instruction Symposium*, volume 2005.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany. Association for Computational Linguistics.

Kai Song, Kun Wang, Heng Yu, Yue Zhang, Zhongqiang Huang, Weihua Luo, Xiangyu Duan, and Min Zhang. 2020. Alignment-enhanced transformer for constraining nmt with pre-specified translations. AAAI.

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.

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.