Towards Precise Lexicon Integration in Neural Machine Translation

Ogün Öz
Cobrainer GmbH
Munich, Germany
ogun.oz@cobrainer.com

Maria Sukhareva
Data Analytics Lab, Siemens AG
Nuremberg, Germany
maria.sukhareva@siemens.com

Abstract

Terminological consistency is an essential requirement for industrial translation. High-quality, hand-crafted terminologies contain entries in their nominal forms. Integrating such a terminology into machine translation is not a trivial task. The MT system must be able to disambiguate homographs on the source side and choose the correct wordform on the target side. In this work, we propose a simple but effective method for homograph disambiguation and a method of wordform selection by introducing multi-choice lexical constraints. We also propose a metric to measure the terminological consistency of the translation. Our results have a significant improvement over the current SOTA in terms of terminological consistency without any loss of the BLEU score. All the code used in this work will be published as open-source.

1 Motivation

The importance of consistent terminology has long been discussed by translation experts (Dagan and Church, 1994; Merkel, 1998; Itagaki et al., 2007; Saraireh, 2001; Byrne, 2006). Terminological standardisation is a critical task for technical and non-technical industrial translation. Patents, technical manuals, and medical instructions rely on consistent usage of technical terminology. But also non-technical news releases, marketing texts, promotion materials, legal and financial documents need to adhere to the same terminology. Byrne (2006) correctly points out that many large companies have their own terminologies that should be used in all texts. Such terminologies prescribe the correct usage of terms and provide not only a list of words that are to be used but also a list of their synonyms that should not be used by writers and translators (so-called negative terms). Sukhareva et al. (2020) describe such terminology for an automotive company and its usage in detail. Not adhering to these rules can be not only confusing for a reader but can also lead to serious legal and financial consequences if it is proven that damage was caused by the ambiguity of the instructions.

Morphologically rich languages also pose a very practical problem for terminology integration: terminological entries are provided in their nominative singular form (Susanto et al., 2020). The SOTA approaches rely on the assumption that the terminological entry can be found as is in the translated text. This is not the case for Slavic languages (e.g. Russian), for which finding the correct wordform on the target side is a key challenge for the terminology integration.

Morphologically poor languages (e.g. English), on the contrary, pose a very different challenge. Homographs appear in such languages not only due to polysemy and homonymy but also due to poor derivational morphology (e.g. a report vs. to report), thus, becoming a very common phenomenon. Liu et al. (2018) show that SOTA neural machine translation (NMT) fails to resolve homography efficiently. Despite being a known issue, the problem has received very little attention from the research community, and we are currently not aware of any prior work that would explicitly address the problem of homographs in the context of terminology integration into machine translation. This paper focuses on the following issues: resolving homographs when the source language is morphologically poor, choosing the right wordform in the morphologically rich target language, and evaluating terminological consistency in the resulting translation. We show that our approach for homograph disambiguation and morphologically flexible lexical constraints significantly improves terminological consistency as compared to the current SOTA.

Proceedings of Recent Advances in Natural Language Processing, pages 1084–1095
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_122

2 Related Work

Previous work can be roughly divided into two groups: approaches that integrate lexicon during inference and approaches that integrate lexicon during training. A constrained decoding approach that has established itself as the SOTA in the past two years is Post and Vilar (2018). They proposed the Dynamic Beam Allocation (DBA) strategy, which decreased the decoding time complexity to constant time in respect to the number of lexical constraints. The proposed algorithm aims to allocate banks dynamically, prioritising the beams that satisfy the most constraints. This algorithm only allows incorporating a single wordform of a constraint, as Dinu et al. (2019) discussed in their work. This is a notable disadvantage of this approach as it assumes an unrealistic precondition that the provided lexical constraints will be correctly inflected. This condition cannot be satisfied when translating into a morphologically rich language.

On the contrary, training time approaches are more flexible in selecting the inflected forms. The SOTA in-training approaches tune a transformer model (Vaswani et al., 2017) towards producing translations that are biased towards an external lexicon. Song et al. (2019) proposed a simple way to copy target side terms into source sentences. Likewise, Dinu et al. (2019) suggested a source sentence modification method by replacing/appending target side terms using additional source factors. Nevertheless, these methods are only encouraging the model to use predefined target terms, whereas constrained decoding methods are enforcing terms' usage. Thus, it can be argued that in-training approaches are inferior to the constrained decoding methods in terms of straightforward terminology integration and, indeed, Dinu et al. (2019) report the terminology usage rate 6-9% less than the constrained decoding method. To ensure the appearance of terms in the output, Michon et al. (2020) use placeholders with the help of morphosyntactic annotations. Even though the approach is effective for choosing a correctly inflected form, it depends on the availability and performance of morphological analysers both in source and target languages.

While all the aforementioned approaches have succeeded in improving the terminological consistency of translations, they essentially rely on a supervised selection of terminological entries. In other words, they assume that the homographs have already been resolved and a correct wordform is provided. Once the discussed approaches are set on a trial under realistic conditions, translation quality deteriorates. Word sense disambiguation is meanwhile a well-researched NLP task, and current state-of-the-art approaches can efficiently resolve homographs (Bohnet et al., 2018; Huang et al., 2019) but due to being time-consuming, are not applicable during translation inference.

3 Data

3.1 Parallel Corpus

For the training of the baseline NMT model, we used preprocessed bilingual WMT18 data[1]. We filtered out sentence pairs that have a length ratio of less than 1/3 or more than 3. We also applied language detection (`langid`) filtering (Lui and Baldwin, 2011) in a tolerant way: The sentence pairs for which `langid` could not predict the expected language in the first 10 predictions are filtered out. Finally, we removed 75,000 sentences with the worst alignment scores (Dyer et al., 2013). All the reported models utilize WordPiece (Wu et al., 2016) for tokenisation. To fine-tune the hyperparameters of the model, we used newstest2014, newstest2018, and newstest2019 as development sets. Newstest2017 is reserved for reporting the results. Since EN → RU newstest2020 was not available during the time of our experiments, we used RU → EN test set including an additional test set (test-ts[2]), as a second set to report the results.

3.2 Terminology Extraction

Despite dictionaries of negative and positive synonyms being standard resources used by industrial translators, they usually cannot be openly shared. Thus, in order to ensure the reproducibility and comparability with previous work, we decided to use openly available resources: WMT Corpus and Russian Wordnet. We believe that such an approximation does not diminish the fairness of the evaluation as we are not focusing on domain adaptation but solely on improving lexical consistency of translation, which is just as applicable to and observable on news translations.

Tab.1 describes the process of generating our pseudo-dictionary of positive and negative terms. The Russian side of the training set is lemmatised

[1] `http://data.statmt.org/wmt18/translation-task/preprocessed/ru-en/`

[2] newstest2020-ruen-src-ts.ru and newstest2020-ruen-ref-ts.en

source	reference
The boat's **engine** had an emergency kill cord.	У **двигателя** лодки был аварийный размыкатель
I opened it up to find out how the **engine** works	я вскрыл его , чтобы проверить , как там работает **мотор**

(a) Sentence pairs where the Russian Wordnet entries двигатель and мотор are aligned to engine.

alignment	occurrence	in synset
engine - двигатель	149	yes
engine - мотор	22	yes
engine - машина	4	no
engine - движущий	3	no
engine - механизм	3	no

(b) Extracted alignments grouped by the source entries. One synset of the most occurring Russian word двигатель is chosen considering the majority vote.

Table 1: The process of generating the terminological dictionary.

and matched against the Russian Wordnet (Chernobay, 2018). We use `fast_align` (Dyer et al., 2013) to extract word alignments of Russian and English sides of the training set. We proceed with finding the English word that is most frequently aligned to all the synonyms in a synset (e.g. "engine" is the most frequent match to "двигатель" *dvigatel'* and "мотор" *motor*). This leaves us with a lexical entry for the English word "engine" and its Russian translations, which are the WordNet synonyms. Finally, we labelled the most frequently aligned Russian synonym in this list as a positive term, and all other Russian synonyms as negative terms (e.g. "двигатель" *dvigatel* is labelled as a positive synonym). Thus, from now on, if an English sentence has a word that occurs in our dictionary, the translator should resort to using the positive term in the translation and avoid negative terms. An example of a terminology entry[3] can be found in Tab. 2.

3.3 Extraction of Wordforms

We further matched the terminology entries in the bilingual training data and kept track of the co-occurrence counts of inflected words to obtain a one-to-many list of wordform candidates per entry. Only the first candidate could be used as a lexical constraint for the related source phrase, whereas all the most frequent k options can be incorporated by our multi-choice lexical constraint approach. In order to extract Russian wordform candidates, we created a list of Russian wordforms most frequently aligned to a single inflected English wordform. As English is a morphologically poor language, we would end up with a list of Russian wordforms that would frequently contain five or more entries. Tab. 3 shows three distinct wordform lists of a terminology entry aligned to an inflected form of the English entry.

[3] https://github.com/term-integration-mt/term-integration-mt

Word	Lang	Usage
engine	en	Positive
двигатель	ru	Positive
мотор	ru	Negative

Table 2: Terminology entry

4 Approach

The approach consists of two major steps. On the source side of the morphologically poor language, it solves the problem of frequent homographs by applying a homograph disambiguator. On the target side of the morphologically rich language, it ensures that the translated term is correctly inflected.

4.1 Homograph Disambiguation for the Morphological Poor Language

Tab. 2 shows an entry in our terminology. All three Russian words are interchangeable synonyms in a certain context. But a straightforward string matching of word *engine* (Tab. 2) with an aim to force the translator to use a certain synonym in the target language would fail: the English word *engine* can also be used in the sense of a search engine (Fig. 1) which would have a Russian literal translation as "search system". In this case, the lexical constraint enforced by our terminology would not be correct

prevails: преобладает, преобладают, преобладать, преобладала
prevailing: преобладающих, преобладающие, преобладает, преобладающее
prevailed: преобладал, преобладали, преобладала, преобладало

Table 3: An example of extracted wordform options depends on the inflections in the source language.

and would cause poor translation quality.

To mitigate this problem, we propose a homograph disambiguation method. Our homograph disambiguation task is simpler than standard word-sense disambiguation (WSD) tasks (e.g. GlossBERT (Huang et al., 2019)) as it suffices to predict whether or not a certain word in the source sentence is used in the same sense as a terminology entry that has the same spelling and, unlike traditional WSD, there is no need to label all the possible senses of this word. We propose a word labelling model, similar to named entity recognition (NER) models, fine-tuned on BERT[4] (Devlin et al., 2019) having only two classes (*T* for Term and *O* for Non-Term). The model tags all the words in a sentence in one forward pass.

In order to create the training data for the homograph disambiguation, we used the same parallel corpus that we used for training the machine translation models. All the training data were processed with a word aligner `fast_align` (Dyer et al., 2013). All the sentences were lemmatised. Every lemma in the Russian sentence was compared against the extracted terminology (Sec. 3.2). If it is found in the terminology as a positive or negative term, we check whether the aligned English lemma is also listed as its translation (Tab. 2). If this is the case, the English word is labelled as "Term", otherwise as "non-Term" (Fig. 1).

The BERT homograph tagger is fine-tuned for 4 epochs on this data.

4.2 Morphology Integration for the Morphologically Rich Language

As described in the Sec. 2, the Dynamic Beam Allocation (DBA)[5] runs in constant time with respect to the number of constraints. The DBA accepts a list of constraint pairs (i.e. a term and its translation). During decoding, the candidates are grouped into banks with the number of banks equal to the number of constraints. If a term is found in the source sentence, then the translation candidates in which term's translation occurs are propagated to a higher bank. The best translation is chosen from the bank with the highest rank (i.e. the ones that have the most satisfied constraints). The drawbacks of this approach is that it matches words without their context and can neither discriminate between homographs (addressed in the previous section) nor choose the correct inflection. As it forces a higher score on the translations that are compliant with the constraint list, the approach is not applicable to translating from a morphologically poor to a morphologically rich language as on one hand there are plenty of homographs on the source side and on the other hand there is a multitude of inflected wordforms on a target side. Constraining a translation on a wrong wordform (e.g., a nominative noun form instead of a dative form) would result in a translator giving a top score to a poor translation.

We propose multi-choice lexical constraints approach that overcomes DBA's limitations and enables the translator to deal with morphologically rich languages by choosing a correct wordform. Similarly to (Post and Vilar, 2018), during inference we allocate candidates to banks. We find the longest possible (in terms of the number of tokens) candidate for every constraint to make sure there will be enough banks for all the possible constraints. Then to prioritise the entirely satisfied constraint phrases regardless of their token count, we rewarded them with the token count of the longest candidate. Without this change, the allocation strategy would be biased towards longer candidates.

Number of constraints The algorithm requires multiple banks to allocate candidate hypotheses. In the worst case, all the longest candidates would need a seat in the bank. For this reason, the number of constraints is the sum of the byte pair encoding (BPE) token counts of the longest constraint options. The size is calculated once since the constraint list remains unchanged during decoding. The number of constraints is calculated as follows:

$$size = \sum_{c \in C} \max_{o} |c_o| \quad (1)$$

where C is the constraint list, and o is a constraint c candidate in multi-choice lexical constraints (MLC) algorithm.

Number of satisfied constraints The satisfied constraint count of hypotheses decides in which bank they should be allocated. The number of banks equals to the maximum possible count if all the longest constraint variants are to be satisfied. However, as the algorithm is biased towards prioritising sentences with the most satisfied constraints, such sentences are longer and have higher overall cross-entropy loss. It causes a significant drop in the general quality of translations, especially if BPE tokenisation is used as more frequent

[4] BERT-Base, Cased (12-layer, 768-Hidden)

[5] For a detailed description of the DBA, refer to Post and Vilar (2018)

Figure 1: Labelling of the training data for homograph disambiguation: English words that are aligned to a synonym in Russian Wordnet synset are labelled as terms, otherwise they are considered to be homographs. ”Двигатель” *”dvigatel’”* and ”мотор” *”motor”* are found in the dictionary, while ”система” *”sistema”* and ”моторный” *”motornyi”* are not.

tokens are usually represented with fewer BPE tokens. To overcome this problem, we calculated the size of the satisfied constraints as follows: given $f(c)$ is the list of the advanced token indices of the constraint c's variant, the number of satisfied constraints in a hypothesis is calculated as:

$$m(c) = \begin{cases} \max_o |c_o|, & \text{if c is entirely met.} \\ \max_o f(c_o), & \text{if c is advanced.} \\ 0, & \text{otherwise.} \end{cases}$$
$$num\_met = \sum_{c \in C} m(c) \quad (2)$$

Set of allowed constraints We keep track of the advanced constraint to make sure we will advance on started but not entirely met constraints. However, when we have multiple variants for a constraint, even if the advanced constraint is known, we might have multiple variants of that constraint as advanced but not fulfilled yet. Therefore, we track the number of advanced tokens for all variants of the constraints. Finally, the set of allowed constraints is defined as the next tokens of all the advanced variants of the advanced constraint. If there is no advanced constraint, the set is simply the initial tokens of all the constraint options. The set $A(C)$ of all the allowed token indices is defined as:

$$A(C) = \begin{cases} f(c_o) + 1, & \exists\ c \text{ with advanced o.} \\ 0 \text{ for all c}, & \text{otherwise.} \end{cases} \quad (3)$$

Advancing on constraints The major difference to the DBA approach is that the advanced constraints have a list of variants on which the algorithm can advance in one step. Therefore, when there is an advanced constraint, all variants are considered as a possible advancement step. For instance, if the initial tokens of the constraint in example (1) are already advanced (пор, ##аж,) in decoding time step t, the algorithm advances on that constraint. The following tokens of both candidates are advanced together for the same hypothesis, which is a usual case when the choices have the same stem, and the only difference is the inflections. Its benefit is not only improving decoding run-time but also distributing the hypotheses more efficiently in the beams.

(1) пор, ##аж, ##ений
пор, ##аж, ##ении

Fig. 2 shows that the run time of the MLC algorithm is comparable with the DBA (Post and Vilar, 2018) in different beam size settings and with different number of wordform choices.

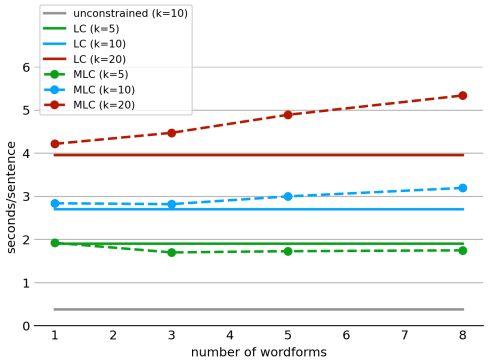

Figure 2: Runtime comparison of (Post and Vilar, 2018) and multi-choice lexical constraints (MLC) as a function of wordform choices per constraints (average runtime per sentences with 2 constraint groups and similar sentence length) where k is beam size.

src.	The **rest** of the people will **rest** until the end of the year.
tr.	**Остальные** люди будут отдыхать до конца года.
ref.	**Остальные** люди отдохнут до конца года.

Table 4: An example sentence pair for terminology usage evaluation.

5 Evaluation

All the models in our experiments were trained in the SOCKEYE[6] toolkit (Hieber et al., 2017). The models that incorporate 6-layer, 8-head transformer architecture are trained 50 epochs on the training corpus (10,402,336 bilingual sentences after preprocessing). We modified the SOCKEYE toolkit to add the multi-choice lexical constraints algorithm and are going to publish the extension as an open-source.

For translation quality evaluation, we report BLEU score (Papineni et al., 2002) using SACREBLEU (Post, 2018),[7] after detokenising the translations. Following Post and Vilar (2018), Dinu et al. (2019), and Susanto et al. (2020), we also report the terminology usage rate to evaluate terminological consistency.

5.1 Terminological F-score

Both BLEU score and terminological usage rate (Post and Vilar, 2018) are not sufficient to evaluate terminological consistency. The usage rate has proven to be seriously flawed as this metric does not account for homographs. Tab. 4 shows an example of a sentence translation that includes a homograph *rest* in its source sentence. Our terminology prescribes translating *rest* as a Russian adjective meaning "remaining" and does not contain an entry that would have the same meaning as its homograph verb *to rest*. The terminology usage rate used in the previous research was calculated in a rather straightforward manner by mere string matching. In our example, it would mean that the metric would only give a perfect score if the verb *rest* was incorrectly translated as its homograph adjective. If this were the case, despite the perfect score, the resulting translation would be of a very poor quality.

As Dougal and Lonsdale (2020) discuss, it is necessary to report an f-score metric when evaluating lexicon injected systems. Their suggested metric TREU intends to mitigate the negative effect of unmatched terminology tokens on BLEU metric assuming the reference sentences do not usually contain terminology promoted tokens. However, to assess the general quality of MT systems clearly, we find it more suitable to use the standard BLEU score. Thus, we require a separate metric based on the precision and recall of the terminology usage.

We propose a terminological f-score to account for precision and recall of the terminology usage in the hypotheses as compared to the reference translations. A similar metric was suggested to evaluate the performance of NMT models for the handling of homographs (Liu et al., 2018). The major difference between our metric and theirs is that we focus on the sense of the word rather than the string by consider all the aligned WordNet synonyms in the reference sentences. The precision and recall per sentence are calculated as follows:

$$P = \sum_{l \in L_S} \frac{\min |l_S|, |l_T|, |l_R|}{|l_T|}$$
$$R = \sum_{l \in L_S} \frac{\min |l_S|, |l_T|, |l_R|}{\min |l_S|, |l_R|} \quad (4)$$

where L_S is the list of the terminological entries that occurred in the source sentence, $|l_S|$ is the occurrence number of terminology entry l in the source sentence, $|l_T|$ is the occurrence number of the positive usage of that entry in the translation sentence, and $|l_R|$ is the occurrence number of both the positive and negative synonyms of the entry l in the reference sentence. Thus, we calculate the precision and recall as 1/1 for the example in Tab. 4, whereas the terminology usage rate is 1/2.

5.2 Quantitative Results

Tab. 5 shows the results of the evaluation in terms of terminology usage rate, terminological f-scores, and BLEU scores for the newstest2017 and newstest2020 testsets. The baseline is a vanilla transformer model trained with the same parameters as all the other models without integrating the terminological dictionary. For the in-training baselines, we reproduce on our data the source-factoring (SF) model with append strategy that was described by Dinu et al. (2019). The inference time baseline is the lexical constraints (LC) approach by Post and

[6] https://github.com/awslabs/sockeye/tree/sockeye_1

[7] The signature is BLEU+case.mixed+lang.en-ru+numrefs.1+smooth.exp+test.wmt17+tok.13a+version.1.4.14

Model	Term. Rate	Term. Prec.	Term. Recall	Term. F1	BLEU (Δ)
baseline	57.43	78.20	81.16	79.65	**33.2**
(Dinu et al., 2019)	81.22	62.76	95.54	75.76	30.2 (-3.0)
SF + BERT	57.17	79.13	81.45	80.27	31.8 (-1.4)
(Post and Vilar, 2018)	99.88	49.04	99.23	65.64	26.0 (-7.2)
MLC	99.68	50.82	99.54	67.29	28.2 (-5.0)
LC + BERT	61.67	75.02	87.30	80.69	31.1 (-2.1)
MLC random	70.71	66.92	86.55	75.48	31.7 (-1.5)
MLC + BERT	61.62	77.35	87.30	**82.03**	32.5 **(-0.7)**

(a) newstest2017

Model	Term. Rate	Term. Prec.	Term. Recall	Term. F1	BLEU (Δ)
baseline	57.33	77.19	75.01	76.08	**28.8**
(Dinu et al., 2019)	81.42	64.72	92.72	76.23	26.4 (-2.4)
SF + BERT	58.27	79.09	77.88	78.48	27.8 (-1.0)
(Post and Vilar, 2018)	99.79	51.13	99.32	67.51	24.6 (-4.2)
MLC	99.51	52.46	99.15	68.62	24.9 (-3.9)
LC + BERT	63.90	74.35	84.73	79.20	27.4 (-1.4)
MLC random	72.31	65.17	82.54	72.83	27.3 (-1.5)
MLC + BERT	63.84	75.84	84.52	**79.94**	28.1 **(-0.7)**

(b) newstest2020 (extracted from ru-en wmt20/test-ts)

Table 5: Terminology usage and BLEU scores of baseline, source factoring by append (SF), lexical constraints (LC) and multi-choice lexical constraints (MLC) (ours) models.

Vilar (2018). We compare the baselines with the following proposed contributions:

1. Introducing homograph disambiguation (+BERT) as described in Sec. 4.1
2. Introducing multi-choice lexical constraints (MLC) for the inference approach as described in Sec. 4.2
3. Combining multi-choice lexical constraints and homograph disambiguation (MLC+BERT)

The evaluation shows that previously proposed SOTA methods for lexica integration by Dinu et al. (2019) and Post and Vilar (2018) suffer from a large decrease in the BLEU score. It also shows that the term usage rate used in the previous research is essentially meaningless for measuring translation quality as even though it has a nearly perfect score for Post and Vilar (2018), the BLEU score greatly dropped. Our approach, on the contrary, showed a significant improvement over all the baselines in terms of terminological f-score without decreasing translation quality. The reasons for the slight decrease of the BLEU score for MLC+BERT are discussed in detail in Sec. 5.3.

5.3 Qualitative Analysis

For a better insight into the results, we manually inspected the Russian translations. One of the primary reasons why MLC+BERT had a slight drop in the BLEU score as compared to the vanilla baseline was that the WMT testset was not tailored to have consistent terminology. We are also not aware of any open-source MT evaluation dataset with terminological consistency in mind. The evaluation showed that this was the reason for the drop in BLEU. Tab. 6 shows translations for which the BLEU score is lower for the MLC+BERT model. This hypothesis was tested by calculating the BLEU score for a subset of test sentences that contain the positive term in the Russian reference translation (80% of newstest2017 and 85% of newstest2017). The results in showed that the difference in the BLEU score between the baseline and our model decreases by more than double if all the test sentences with negative terms are eliminated (see Appendix A).

As compared to other baselines, our method greatly improves the quality of the translation for Post and Vilar (2018) and Dinu et al. (2019). Post and Vilar (2018) baseline is particularly prone to hallucinate Lee et al. (2018) if a lexical constraint

EN	Kvyat parked his **car** in one of the safety zones.	**Terminology**
RU	Квят припарковал **машину** в одной из зон безопасности.	car
baseline	Квят припарковал свою **машину** в одной из зон безопасности.	автомобиль (pos)
MLC+BERT	Квят припарковал свой **автомобиль** в одной из зон безопасности.	машина (neg)

Table 6: An example from the newstest2020 evaluation set. The Russian gold sentence and the baseline contain a negative term. The MLC+BERT translation uses a positive interchangeable syllable. Even though the translation is perfectly fine, the BLEU score is lower for MLC+BERT.

is a homograph or is not correctly inflected (see Appendix B). In this case, the model generates an output till it reaches the maximum length. For example, the output of the LC baseline has 8% more characters than the reference translations. In comparison, the vanilla baseline has only 0.5% more characters and the MLC+BERT has exactly the same amount of characters. The manual evaluation showed that reducing hallucinations is the reason for the large increase of the BLEU as compared to the SF and LC baselines.

We also examined the effect of automatically generated lexicon on the translation quality. While we found cases in which positive terms were not perfect synonyms and were not interchangeable with negative terms, the homograph disambiguation seemed to show certain robustness by labelling the English term only if they occurred in the context that was common for negative and positive Russian translations. While we still believe that better results could be achieved in real-life settings where a high-quality dictionary would be used, our examination showed that there was no unreasonable error propagation from the usage of an automatically extracted dictionary.

The greatest weakness that we found during qualitative examination lies in how the top inflected candidates are scored in MLC. The MLC model takes a list of top *n* Russian wordforms that are most frequently aligned to a given English wordform of a term. In rare cases, an acceptable wordform does not appear to be in the top *n* list. In this case, the translation ends up being grammatically incorrect or hallucinates in a similar sense as the LC baseline. A possible solution for this would be generating the top *n* choices for MLC in a more elaborated manner e.g. by considering the position in the sentence or even using syntactic information. For now, we leave exploring those options for future work.

5.4 Evaluation of Homograph Disambiguation

The homograph disambiguator was trained on artificially created labels, and we are not in possession of any gold standard data for the direct evaluation. We assume that evaluating the approach on the artificially labelled data will not ensure the objectivity of such an evaluation and both train and testset will contain the same errors. For transparency, we still provide the scores in Appendix (Tab. 8).

Thus, measuring the effect of homograph disambiguator on the downstream translation task is more sound. To make sure that the improvement of the terminological f-score is caused by the homograph disambiguation and not by the reduction of the number of lexical constraints, we introduce the $\text{MLC}_{\text{random}}$ baseline (see Tab. 5). We have calculated the total amount of constrained terms after applying the homograph disambiguation (+BERT) and randomly labelled the same amount of terms to be constrained in the original testsets. The evaluation results showed that the f-score dropped by 7% for the randomly labelled dataset, thus, proving that our homograph disambiguation is the actual cause of the f-score's improvement.

5.5 Runtime Analysis

In order to ensure that MLC is also feasible for real-life usage, we compared the inference speed between the Post and Vilar (2018) and our MLC input (Fig. 2). As well as the DBA algorithm, MLC makes sure that the number of hypotheses is limited by the beam size. Thus, the runtime complexity of our approach is constant in the number of constraints. We have made an interesting observation that MLC is actually faster than LC for the beam size of 5 and slightly slows down for the beam size of 10. We have found the following explanation for such behaviour: Lexical constraints expect a large beam size in order to be able to generate enough hypotheses with the provided lexical constraints. The DBA does not allow a beam to generate the end of sentence symbol unless the constraints are met. Once a translation is incorrectly constrained on a homograph or on a wordform that cannot occur in translation, the beam cannot terminate unless it reaches the maximum length, and, thus, it negatively influences the inference time. On the con-

trary, the MLC allows a beam to terminate which makes it more time efficient.

6 Conclusion

We have presented an approach for terminology integration into a neural machine translation from a morphologically poor into a morphologically rich language. Our work makes the following contributions:

1. Disambiguation of the homographs in the morphologically poor language.
2. Multi-choice lexical constraints to ensure the correct choice of an inflected target wordform in the morphologically rich language.
3. A metric that takes into account precision and recall of terminology usage.

We propose a solution to the problem of rich morphology in the target language by presenting multi-choice lexical constraints and show that our combined approach (MLC+BERT) has a significantly[8] better f-score than all the other models.

References

Bernd Bohnet, Ryan McDonald, Gonçalo Simões, Daniel Andor, Emily Pitler, and Joshua Maynez. 2018. Morphosyntactic tagging with a meta-BiLSTM model over context sensitive token encodings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2642–2652, Melbourne, Australia. Association for Computational Linguistics.

J. Byrne. 2006. *Technical Translation: Usability Strategies for Translating Technical Documentation*. Springer Netherlands.

Yuliya Chernobay. 2018. Building wordnet for russian language from ru.wiktionary. In *Artificial Intelligence and Natural Language*, pages 113–120, Cham. Springer International Publishing.

Ido Dagan and Kenneth Church. 1994. Termight: Identifying and translating technical terminology. In *Fourth Conference on Applied Natural Language Processing*, pages 34–40.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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, Atlanta, Georgia. Association for Computational Linguistics.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *arXiv preprint arXiv:1712.05690*.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. 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 3509–3514, Hong Kong, China. Association for Computational Linguistics.

Masaki Itagaki, Takako Aikawa, and Xiaodong He. 2007. Automatic validation of terminology translation consistency with statistical method. *Proceedings of MT summit XI*, pages 269–274.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation.

Frederick Liu, Han Lu, and Graham Neubig. 2018. Handling homographs in 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 1336–1345, New Orleans, Louisiana. Association for Computational Linguistics.

Marco Lui and Timothy Baldwin. 2011. Cross-domain feature selection for language identification. In *Proceedings of 5th International Joint Conference on*

[8] We used McNemar's significance test (McNemar, 1947). The significant difference is defined as $p < 0.05$.

Natural Language Processing, pages 553–561, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Magnus Merkel. 1998. Consistency and variation in technical translation: A study of translators' attitudes. *Unity in diversity*, pages 137–149.

Elise Michon, Josep Crego, and Jean Senellart. 2020. Integrating domain terminology into neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3925–3937, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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, Philadelphia, Pennsylvania, USA. 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.

Muhammad A Saraireh. 2001. Inconsistency in technical terminology: A problem for standardization in arabic. *Babel*, 47(1):10–21.

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.

Maria Sukhareva, Olgierd Grodzki, and Bernhard Pflugfelder. 2020. Industrial machine translation system for automotive domain. In *Proceedings of the LREC2020 Industry Track*, pages 31–35, Marseille, France. European Language Resources Association (ELRA).

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

A Further Quantitative Analysis

A.1 BLEU Scores of Filtered Datasets

Model	BLEU (Δ)
baseline	**33.7**
SF + BERT	32.3 (-1.4)
MLC + BERT	33.2 **(-0.5)**

(a) newstest2017

Model	BLEU (Δ)
baseline	**29.6**
SF + BERT	29.1 (-0.5)
MLC + BERT	29.3 **(-0.3)**

(b) newstest2020

Table 7: BLEU scores after filtering the sentences having at least one negative term.

A.2 Direct Evaluation of Homograph Disambiguation

testset	precision	recall	f-score
newstest2017	76.17	56.64	64.97
newstest2020	66.92	79.90	72.84

Table 8: Evaluation table for homograph disambiguation task. Since there is no gold labels, predicted labels are compared against the artificially created labels.

B Qualitative Comparison of Translation Systems

Type of error	Hallucination after correct translation
EN	It was earlier reported that capital CSKA player Konstantin Kuchaev spoke out against the introduction of VAR (Video Assistant Referee).
RU	Ранее сообщалось, что футболист столичного ЦСКА Константин Кучаев высказался против внедрения VAR.
baseline	Ранее сообщалось, что столичный игрок ЦСКА Константин Кучаев выступил против введения ВАР (видео помощника судьи).
LC	Ранее сообщалось, что столичный игрок ЦСКА Константин Кучаев выступил против введения ВАР (видео помощника судьи*, который говорил о том, что арбитру докладе не удалось выступить с рефери.*
MLC+BERT	Ранее сообщалось, что столичный игрок ЦСКА Константин Кучаев выступил против введения ВАР (видео помощника судьи).
Comment	The LC model generates a string after comma (marked in italics) that does not occur in the source text nor meaningful in the context. It happens because the lexicon prescribes to translate ”report” as a noun meaning ”an account given of a particular matter” *доклад*, while the source actually has a homograph verb ”to report”. The LC model generates a correct translation and proceeds to hallucinate till it finally produces a sentence with ”a report”. It leads to not only longer nonsensical output but also to longer inference time. The homograph disambiguation (MLC + BERT) correctly marks ”report” as a non-term, thus, preventing the model to force a constraint on this sentence
Type of error	Hallucination with a grammatically correct sentence
EN	As reported by Chempionat, the 41-year-old specialist flew into Moscow to weigh up the possibility of working at one of Russia’s clubs.
RU	Как сообщает ”Чемпионат”, 41-летний специалист прилетел в Москву, чтобы изучить возможность найти работу в каком-нибудь российском клубе.
LC	Как сообщает Chempionat, 41-старый специалист вылетел в Москву, чтобы в докладе проанализировать возможность работы в одном из российских клубов.
MLC+BERT	Как сообщает Chempionat, 41-летний специалист вылетел в Москву, чтобы взвесить возможность работы в одном из российских клубов.
Comment	As in the previous example, the LC model forces to use the homograph noun ”a report” to be a translation of the verb ”to report”. Unlike the example above, the model does not produce a correct translation at any point and generates a sentence with an entirely different meaning: ”As reported by Chempionat, the 41-year old specialist got on a flight to Moscow to analyse in his report possibilities of working at one of Russia’s clubs.” This kind of translations are particularly dangerous, as it would be extremely difficult for a native speaker without looking at the source to detect that the translation completely fails to convey the meaning. The homograph disambiguation solves this problem and the translation is correct.
Type of error	Hallucination with an ungrammatical sentence
EN	Documents obtained by the publication, reveal that the owners of TikTok (ByteDance company) with the help of their app are promoting Chinese **foreign** policy goals overseas.

RU	В документах, оказавшихся у издания, рассказывается, что владелец TikTok (компания ByteDance) с помощью приложения продвигает цели внешней политики Китая за рубежом.
baseline	Документы, полученные публикацией, показывают, что владельцы TikTok (ByteDance company) с помощью своего приложения продвигают китайские внешнеполитические цели за рубежом.
LC	Зарубежных иностранных владельцев помочь способствовать показать целей компании TikTok (ByteDance company), полученную в результате публикации, в приложении.
MLC+BERT	Документы, полученные изданием, показывают, что владельцы компании TikTok (ByteDance) с помощью своего приложения продвигают китайские цели внешней политики за рубежом.
Comment	The LC baseline forces to translate **foreign** as *иностранный* which is not applicable in this context. The LC baseline generates a nonsensical sequence of words. This type of error is less harmful that the one described above as a native speaker can immediately spot that translation is incorrect. The MLC+BERT solves this problem and the translation is correct.
Type of error	A wrong wordform as lexical constraint
EN	Documents obtained by the publication, reveal that the owners of TikTok (ByteDance company) with the help of their app are promoting Chinese **foreign** policy goals overseas.
RU	В документах, оказавшихся у издания, рассказывается, что владелец TikTok (компания ByteDance) с помощью приложения продвигает цели внешней политики Китая за рубежом.
baseline	Документы, полученные публикацией, показывают, что владельцы TikTok (ByteDance company) с помощью своего приложения продвигают китайские внешнеполитические цели за рубежом.
LC+BERT	Документы, полученные изданием, показывают, что **владельцев** компании TikTok (ByteDance) с помощью своего приложения продвигают китайские внешнеполитические целей за рубежом.
MLC+BERT	Документы, полученные изданием, показывают, что **владельцы** компании TikTok (ByteDance) с помощью своего приложения продвигают китайские цели внешней политики за рубежом.
Comment	The error described in the previous example was resolved by the homograph disambiguation. However, the LC + BERT model produced a grammatically incorrect translation as the constraint for word ”owners” was given in a wrongly inflected form of Genitive plural **владельцев** . The MLC+BERT solves this problem by providing a list of inflected forms and the result is a correct translation of the word in Nominative plural. Interestingly, the reference translation is incorrect and translates ”owners” as singular nominative ”owner”. **владельцы**.
Type of error	Inconsistent terminology usage in the test set
EN	Roman Zaripov, founder of the Our Digital agency, agreed with Bogdanov: ”The **main** rules for TikTok users are listed in the user agreement: no posting shocking content, discriminatory rhetoric and so on.”
RU	С Богдановым соглашается основатель агентства Our Didgital Роман Зарипов: ”**Основные** правила для пользователей TikTok перечисляет в пользовательском соглашении: нельзя выкладывать шокирующий контент, дискриминирующие высказывания и так далее”.
baseline	Роман Зарипов, основатель нашего цифрового агентства, согласился с Богдановым : ”**основные** правила для пользователей TikTok перечислены в пользовательском соглашении: никакого размещения шокирующего контента, дискриминационной риторики и так далее”.
MLC+BERT	Роман Зарипов, основатель нашего цифрового агентства, согласился с Богдановым : ”**главными** правилами для пользователей TikTok являются пользовательские соглашения: никакого размещения шокирующего контента, дискриминационной риторики и т.д”..
Comment	Both baseline and MLC + BERT produced correct translations. Word ”main” is prescribed to be translated as **главный** by our terminology. However, in the baseline it is translated with a negative term **основной** while both translations are correct, the BLEU score for our model will be penalized for using a synonym of a word used in the reference translation.
Type of error	Insufficient coverage by the lexicon
EN	This historic trajectory **cannot** be stopped by anyone or any force, said Xiaoguang.
RU	Эта историческая тенденция **не может быть** остановлена никем и никакими силами, подчеркнул Ма Сяогуан.
baseline	Эта историческая траектория **не может быть** остановлена ни кем или какой-либо силой , сказал Сяоугуань.
MLC+BERT	Эту историческую траекторию **нельзя** остановить никем или какой-либо силой, сказал Сяоугуань.
Comment	The lexicon only includes **нельзя** as a positive term and **невозможно** as a negative term. The Russian phrase **не может быть** is a valid translation but was not included in the Russian WordNet. While the homograph disambiguator correctly labelled the ”cannot” as a term, it was not labelled as a positive term in the test data as neither positive nor negative term was aligned to it. This is a reason why we believe that the evaluation against the random baseline MLC+BERT$_{\text{random}}$ (Tab. 5) is more reliable than a mere f-score on the test set.

Table 9: Examples of various errors that were identified during qualitative analysis