Investigating automatic and manual filtering methods to produce MT-ready glossaries from existing ones

Maria Afara
AI R&D team, Acolad
mafara@acolad.com

Randy Scansani
AI R&D team, Acolad
rscansani@acolad.com

Loïc Dugast
AI R&D team, Acolad
ldugast@acolad.com

Abstract

Commercial Machine Translation (MT) providers offer functionalities that allow users to leverage bilingual glossaries. This poses the question of how to turn glossaries that were intended to be used by a human translator into MT-ready ones, removing entries that could harm the MT output. We present two automatic filtering approaches – one based on rules and the second one relying on a translation memory – and a manual filtering procedure carried out by a linguist. The resulting glossaries are added to an MT model. The outputs are compared against a baseline where no glossary is used and an output produced using the original glossary. The present work aims at investigating if any of these filtering methods can bring a higher terminology accuracy without negative effects on the overall quality. Results are measured with terminology accuracy and Translation Edit Rate. We test our filters on two language pairs, En–Fr and De–En. Results show that some of the automatically filtered glossaries may help reach a better balance between accuracy and overall quality, replacing the costly manual process.

1 Introduction

The ability to correctly and consistently translate domain-specific or customer-specific terminology is key in the field of translation. To accommodate for this need, Machine Translation (MT) providers have started to offer terminology features that enforce glossary entries at runtime.[1] The availability of such features can be particularly advantageous for Language Service Providers (LSPs), giving them an opportunity to offer a terminology accurate MT output in a scenario in which training a model from scratch is not an option.

However, such glossaries were created to be used by human translators, relying on their ability to, e.g., disambiguate terms before inserting them in the target text. Also, glossaries are often created by customers without the help of terminologists. As a result, they might not be ready to be used by MT, since they might contain entries that harm the output quality (Bergmanis et al., 2021; Guerrero, 2020; Scansani and Dugast, 2021).

The creation of a pipeline to clean glossaries can help MT users leverage their terminology data base. The pipeline can be based on a manual intervention, which can be time-consuming, or rely on an automatic procedure. Either way, two operations may be involved, i.e. removing entries that are not helpful and/or editing them.

Automatically editing entries is not a trivial task. For example, automatically editing term entries where several term alternatives are separated by slashes poses the question of which alternative(s) to keep. Also, in some cases the slash is used to separate parts of a compound or of a multi-word term (e.g. the German term “Abluft-/Motorfilter” should be split into “Abluftfilter” and “Motorfilter”). Editing terms with parentheses is not trivial either. In some cases what is inside the parentheses is part of the term, e.g. the German term “Länge Base” translated as “Length (base)”. In other cases

[1] Two examples of glossary functionalities are https://bit.ly/2U5os9v and https://bit.ly/3H4x4zy.

Domain	Lang. pair	Sent. pairs	Term pairs	
			Original	Validated
Electrical devices	DE>EN	1,725	3,050	1,898
	EN>DE	1,698		
Sportswear	EN>FR	1,951	1,758	1,190
	FR>EN	1,544		

Table 1: Number of sentence pairs in the test set, and number of glossary entries in the original glossary and in the manually validated one for each of the four use cases tested.

it is not, and it should be removed, e.g. "Kühlung (z. B. von Notebooks)" translated as "cooling", where the content of the parentheses provide context for the term. For these reasons, we will rather focus on filtering out such invalid entries (more example provided in Sect. 3.3).

In this paper we present procedures to filter glossaries automatically and manually. We investigate the results each glossary yields in terms of terminology accuracy and overall output quality – as measured by automatic metrics – when it is leveraged by the glossary feature of commercial neural MT (NMT) providers. Our main contribution is to investigate if any of these filtering methods brings improvements to terminology accuracy with respect to the baseline, without worsening the overall quality compared to the output where the whole glossary is used. Ideally, a better terminology accuracy should bring a higher overall quality, but since not much is known about how MT providers implement their glossary feature, we also want to check if this feature introduces side effects. The results obtained with the filtered glossaries are compared to those obtained when no glossary is used and when the original one is applied.

Two automatic glossary cleaning techniques are presented (see Sect. 3.3). One is based on rules to remove noisy entries. The second one also leverages a Translation Memory (TM) to remove entries that are not used consistently in the translated contents. Both filtering techniques are applied to two use cases, i.e. *Sport equipment* English–French and *Electrical devices* German–English. Two providers are tested and their performance is evaluated based on terminology accuracy and overall output quality (see Sects. 3.2 and 3.5). Some sentences are then manually inspected to highlight interesting patterns in the outputs.

The remainder of the paper is structured as follows. Section 2 provides a brief overview of the literature in the field of terminology and NMT. The experimental setup (data sets, MT providers, evaluation method, filtering methods and experiments carried out) is outlined in Section 3 and its subsections, and the results are presented in Section 4. Section 4.3 offers a review of some examples. Results are then discussed in Section 5, together with suggestions on future work.

2 Background

Several different approaches have been developed to enforce glossary terms in the NMT output. A growing interest in this field is testified by the first Shared Task on Machine Translation Using Terminologies in the framework of WMT 2021 (Alam et al., 2021b). The methods developed so far can be broadly grouped in two categories. Some of them are based on the idea of injecting terms from a glossary into the MT output as constraints posed at decoding time (Chatterjee et al., 2017; Dougal and Lonsdale, 2020; Hasler et al., 2018; Hokamp and Liu, 2017). Other works build on the idea of adding *soft* constraints by annotating the source side of the training data (Ailem et al., 2021; Bergmanis and Pinnis, 2021a; Bergmanis and Pinnis, 2021b; Dinu et al., 2019; Exel et al., 2020).

Commercial MT providers do not disclose how their glossary feature is implemented, thus we do not now if they apply one of the approaches mentioned so far, and little work has investigated the performance of commercial models when enhanced by a glossary. Guerrero (2020) compared the work of translators post-editing the output with and without the glossary. Scansani and Dugast (2021) have investigated how the performance of a number of MT models changes when a preexisting glossary is added. Both works conclude that pre-existing glossaries should be filtered before being used for MT. The need of preparing glossaries so that they are MT-ready is also underlined by Bergmanis *et al.* (2021). However, to the best of our knowledge, the paper by Bergmanis and Pinnis (2021a) is the only one to have compared the impact of different glossary filtering approaches on the MT output. In their work, noisy and inconsistent entries are automatically filtered

Issues	Rule-based filter decision	TM-based filter decision
Duplicates	Keep first	Keep first
TB inconsistent	Keep first	Keep first
Format issues	Discard	Discard if no match in the TM
TM inconsistent	*na*	Discard based on TM inconsistency
TM unmatched	*na*	Discard

Table 2: Table summarizing the decisions taken for each issue found in the TB by the Rule-based filter and by the TM-based filter.

Issues	Manual filter decision
TB inconsistent	Keep first
Format issues	Edit/Discard
Wrong translation	Discard
Invalid term	Discard
Typo/misspelling	Edit
Contains term info	Edit
Contains alternatives	Edit

Table 3: Table summarizing the decisions taken for each issue found in the termbase (TB) in the manually filtered glossary.

out – in some cases with the help of word alignment. In the present work, we test the use of a TM to validate or discard glossary entries and we compare automatic filtering to the manual procedure.

3 Experimental setup

3.1 Data set

Two different data sets belonging to two domains are used for our experiments. One domain is *Electrical devices* and the language combination is German–English. The other domain is *Sport equipment* and the language combination is English–French. This allows us to run the tests on different content types, and on different language pairs, where at least one (En–Fr) is not into English and the ability to handle term inflections is therefore more relevant. Number of sentences and of term pairs used is displayed in Table 1.

For each use case, a pre-existing glossary was manually validated by a linguist specialized in the domain. During the validation procedure – explained in Sect. 3.4 – the linguist could validate, remove or edit terms. The validated glossary – composed of the validated entries only – was then used for the terminology accuracy evaluation.

The test set was created by extracting sentences from a bilingual corpus that had at least one source match from the following terms: terms in the original glossary that were validated by the linguist, terms in the original glossary that were removed by the linguist, and terms in the original glossary that were edited by the linguist. In this last case, we look for matches of the edited version of the term rather than the original one.

By including both validated and unvalidated/edited terms, we test for two distinct cases. Sentences with matches from validated entries are the ones where we expect any glossary to have a positive impact on accuracy and output quality, unless the filter is erroneously removing valid entries. The second case is that of sentences with matches from unvalidated/edited terms, i.e. sentences where we expect a glossary to have a positive impact only if the glossary was filtered, and if the filter removed such invalid terms.

3.2 MT Providers

We chose two NMT providers whose glossary feature implementations differ in terms of source term matching and target term insertion. Although no specific information is offered by the providers, preliminary tests we carried out showed that Provider 1 is able to inflect terms so that their morphological form fits the rest of the sentence, whereas Provider 2 enforces terms in the output without any adjustment. Regarding source term matching, Provider 1 is able to match terms on a lemma level and regardless of their casing. Provider 2 matches terms only if the term in the source sentence has the same casing and the same morphological form as the term in the glossary.

We chose not to reveal the name of the providers used because we are not aiming at benchmarking them, but rather at focusing on the results we get with our filtering approaches.

3.3 Automatic filtering methods

Two filtering methods are used and tested. One relies on rules to remove noisy entries. The second one is based on the same rules as the first, but it leverages a TM to confirm or deny the rule-based decision. If the rules identify an entry as noisy, but it has matches in the TM, the TM-based filter retains it while the rule-based filter discards it. Additionally, the TM-based filter removes entries that are not used consistently or at all in the translated contents. The rules were mainly decided based on the issues observed in a number of Termbases (TB), but also based on the suggestions set out in

Bergmanis *et al.* (2021). Table 2 summarizes the different filter decisions for each of the two methods. More information on the issues follow.

Duplicates: Usually MT providers require to use glossaries that do not contain source-target duplicates. When a term pair is duplicated, we always keep only one.

TB inconsistent: Glossaries are usually expected to contain only one single instance of each source term and just one translation. This is especially key for MT. An MT engine cannot know which target term to pick in case of inconsistencies in the glossary, which might lead to inconsistent translations. Given a source term which has inconsistent translations in a TB, we keep the first entry occurring in the TB.

Has format issues: The following entries are automatically discarded by the rule-based decision filter. In the case of the TM-based filter they are kept if they have matches in the TM.

- Extra white spaces
- Numbers: dates, numbered paragraph titles, etc., e.g. "1 from 08/1992 to 09/2001" "or 2 - Type of Product Range".
- Punctuation: slashes, pipes, brackets and others are sometimes added to the term, especially to separate term alternatives – e.g. "Screw / Dowel / Nut" – or when explanations and domain/contextual information are added to the term field – "expose <photo>", "bottom (of a bag)", "Tasche|Case".

The following term pairs are filtered out only by the TM-based filter:

TM inconsistent: When a source term occurring in the glossary is translated inconsistently in the TM, it might mean that the glossary entry is not correct and/or that the translator did not enforce it, or that the entry was added to the glossary at a later stage. We therefore remove such entries based on different thresholds (see Table 4). A 40% threshold means that a glossary entry is kept if its source-target matches in the TM correspond at least to 40% of its total number of source matches. When the percentage increases, more terms are removed.

TM unmatched: Term pairs that are not matching in the TM are removed, based on the assumption that if translators are not inserting them, they might not be relevant for the domain or may even harm the output.

3.4 Manual filtering method

Each glossary was validated by one linguist specialized in the domain. The linguists were provided with instructions on how to clean terms, and also with general information on the use of terminology for MT. Guidelines did not include any specific information on the NMT providers used, so that the validation process was not biased towards the terminology injection approach of a provider. They were asked to label each entry as: *to be kept*, *to be removed*, or *edited*. In the last two cases, a reason had to be picked among those provided (e.g. long term, duplicate source term, punctuation in the term field, wrong translation, etc.). In case a term was labelled as *edited*, a new, correct version of the term had to be provided by the linguist. As introduced in Sect. 1, the present work focuses on methods to filter out terms. However, the manual process included the edition of some terms, which gave us the possibility to have a correct version of some of the invalid terms. In the scope of the present work, the edited terms are used only to produce the test set (see Sect. 3.1). Instead, the subset containing the validated terms only is used to compute the accuracy (see Sect. 4) and was leveraged by the MT providers in the *manual filter* experiments. We acknowledge that using the same glossary in one of the experiments and in the evaluation is a limitation of this work. However, the evaluation should be carried out using a manually validated glossary, which left us without other viable options than using this glossary for the evaluation as well.

More information on the issues in Table 3 that were not already described in Sect. 3.3 follows.

Wrong translation: One term in the entry is valid, but its translation is not correct.

Invalid term: (one of) the terms in the entry do not comply with the standard definition of term[2].

Term info in term field: In some cases, the term field of the glossary contains information on the domain of a term, e.g. "exposure (photography)". Such piece of information is erroneously added to the term field as an extra information for the translator. In the automatic filtering, this is handled by removing entries containing punctuation. In the manual filter, we ask the linguist to correct

[2] "A term is a graphic and/or phonic sign - a word or group of words, a compound word or a locution, an abbreviation - that allows to express a special concept related to concrete or abstract objects [...] that can be uniquely defined within a given discipline." (Riediger, 2018, our translation).

Electrical devices DE>EN	**Provider 1**		**Provider 2**		**Glossary size**
	TER ↓	**Acc. ↑**	**TER ↓**	**Acc. ↑**	
Baseline	**26.7**	79.8	**27.6**	82.8	0
Whole glossary	29.7	**96.7**	30.6	96	3033
Rule-based filter	29.6	**96.7**	30.4	96.1	2963
Manual filter	28.7	96.6	29.9	**96.3**	1590
TM-based filter > 40%	28.6	92.4	30	92	2188
TM-based filter > 60%	28.2	90.5	29.8	89.6	2097
TM-based filter > 80%	27.9	88.3	29.5	88.1	2007
TM-based filter > 90%	27.8	86.9	29.4	87	1949
TM-based filter 100%	27.2	82.0	28.2	83.58	1852
Sportswear EN>FR	**Provider 1**		**Provider 2**		**Glossary size**
	TER ↓	**Acc. ↑**	**TER ↓**	**Acc. ↑**	
Baseline	60.5	37.4	60.4	38.3	0
Whole glossary	58.1	70.2	59.5	**63.8**	1734
Rule-based filter	58.1	70	59.5	63.4	1527
Manual filter	58.9	**71**	59.6	63.7	1190
TM-based filter > 40%	**57.6**	68.4	**59.1**	60.7	915
TM-based filter > 60%	58	62.2	59.8	51.5	762
TM-based filter > 80%	58.7	50.5	59.8	48.8	697
TM-based filter > 90%	59.7	44.1	60.2	43.1	631
TM-based filter 100%	60.3	38	60.28	39.3	577

Table 4: TER and accuracy results for Electrical devices De–En and Sportswear En–Fr, for both providers tested. The rightmost column contains the total number of entries in each glossary. Each row represents one of the filtering methods applied.

such entries by removing the information.

Term contains alternatives: Some term entries contain more than one term separated, e.g., by pipes or slashes (see examples in Sect. 1). The linguist was asked to keep the best alternative based on his/her knowledge of the text domain and remove the other ones. In the automatic filtering, this is handled by removing entries containing punctuation.

3.5 Evaluation metrics and method

The assessment of the MT output aims at investigating its overall quality and its terminology consistency, comparing a baseline (no glossary is added) to the outputs obtained using the whole glossary, the automatically filtered ones, and the manually filtered one. Translation Edit Rate (TER) (Snover et al., 2006), case insensitive, is used as quality metric, whereas glossary compliance is measured by terminology accuracy, as suggested in Alam *et al.* (2021a). To compute accuracy, we look for occurrences of glossary term pairs in the source-target text. Both the text and the glossary are lemmatized and lowercased. In case of overlapping matches, we keep the longest matching entry only. Accuracy is then computed as the proportion between glossary matches in the source text and source-target glossary matches.

The first step of our evaluation process is to compute accuracy and TER on the whole data set (Sect. 4.1). In order to have a better understanding of how the usage of glossaries impacts the output quality, we then perform a sentence-level analysis (Sect. 4.2). Indeed, a minor TER or accuracy variation on the whole data set may hide, e.g., a high number of small differences between a sentence translated without the glossary *vs.* a sentence translated with the glossary, or a low number of sentences with large differences.

4 Experimental results

4.1 TER and accuracy on the whole data sets

Results in Table 4 show that the filtering approaches have a different impact on the output based on the use case and on the provider. Also, it shows that the best accuracy results (in bold in the *Accuracy* columns) do not correspond to the best TER score (in bold in the TER column).

Electrical devices De–En. In this use case, the baseline has the best TER score (26.7% with Provider 1 and 27.6% with Provider 2). As expected, adding a glossary always improves accuracy with respect to the baseline. The whole glossary, the rule-based filter and the manual filter achieve the highest accuracy scores for both providers, ranging from 96.1% to 96.7%. However, they also have the worst TER scores for Provider 1, while TER results for Provider 2 are less clear-cut, and all filtering methods – excluding the best one – range from 30.6% to 29.4%. The high accuracy achieved by the baseline (79.8%

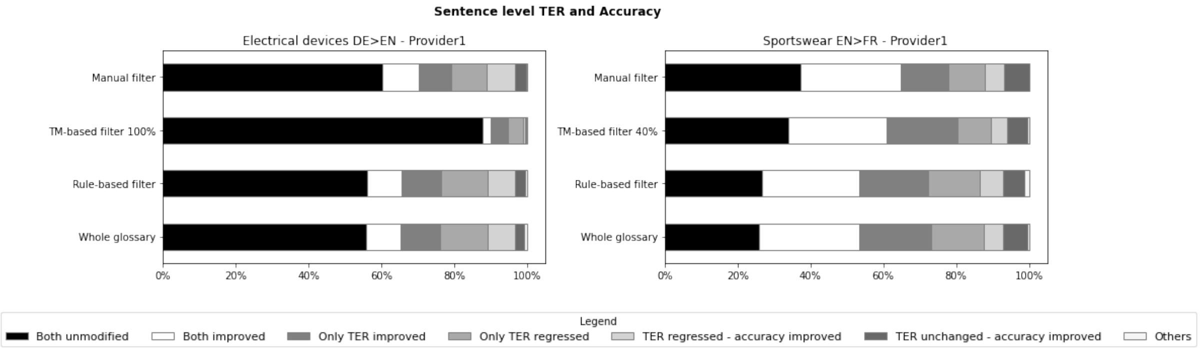

Figure 1: For each use case, we report on the percentage of sentences produced by Provider 1 that were assigned to one of the seven categories in the legend. The categories refer to the comparison between the baseline and the output produced with the filtered glossaries.

and 82.8%) suggests that the terminology for this use case is not highly specific and a generic model without any glossary attached to it can already handle most of the terms correctly. If the terms are rather generic, some of them might have different translations depending on the context, and enforcing them might harm the quality. Indeed, for De–En in general, achieving a very high accuracy is not possible without hampering the overall quality. The TM-based filter with varying thresholds (see Sect. 3.3) shows that, for Provider 1, a less restrictive threshold (e.g. >40%) leads to a high accuracy, which in turn causes a slight TER increase. The most restrictive filter (100% threshold) reaches an 82% accuracy and a 27.2% TER, the best TER obtained with a glossary – and the closest TER score to that of the baseline. A similar pattern is shown by Provider 2. The manually validated glossary is not outperforming the automatically filtered ones neither in terms of TER, nor in terms of accuracy, where it slightly outperformed the rule-based glossary. In general, there seems to be an accuracy cutoff over which TER cannot improve.

Sportswear En–Fr. In this use case, the highest accuracy is achieved by the manual filter for Provider 1 (71%), and by the whole glossary for Provider 2 (63.8%). The best TER is obtained with TM-based filter > 40% (57.6% and 59.1%). The very low accuracy obtained by the baseline suggests that the term entries in this glossary are highly domain-specific, and based on the TER results the general quality is benefiting from the use of a glossary, which was not the case for Electrical devices. Again, the filters show that aiming at the highest possible accuracy brings a lower TER. This is especially true for Provider 1, where the TM-based filter 100% has a 38% accuracy and a 60.3% TER, whereas the less restrictive 40% threshold increases accuracy to 68.4%, with a 57.6% TER. Similarly to the previous use case, the manually filtered glossary, although yielding a good accuracy score, is not outperforming the automatically filtered glossaries in terms of TER. To conclude, Sportswear results seem to show that filtering the glossary brings improvements, although small, with respect to using a whole glossary.

4.2 Sentence-level analysis

To gain a better understanding of how TER and accuracy are changing, we carried out a sentence-level analysis that compares the output of the baseline against the output obtained with each of the other glossaries. For the TM-based filter, we picked the one with the 100% threshold for Electrical Devices and the one with a 40% threshold for Sportswear. We limit the scope of this analysis to the results obtained with Provider 1, which implements a more sophisticated terminology feature than Provider 2 (see Sect. 3.2).

Output sentences were grouped in seven categories (see legend in Fig. 1), based on TER and accuracy differences with respect to the baseline. For example, *Both improved* includes all sentences where both TER and accuracy are better in that specific output than in the baseline.

Fig. 1, shows that more than half of the sentences are the same as in the baseline in all outputs of Electrical devices. The impact of the glossary is thus limited to the remaining, smaller portion. As for Sportswear, only a small part of the outputs stayed unchanged. Regardless of the filtering method, for both use cases, a good number of sentences sees changes in TER (see *Only TER regressed* and *Only TER improved* categories)

		Sentence	TER	Acc
1	Source	Men's Short Sleeve **Baselayer**		
	Reference	*Première couche* à manches courtes pour homme		
	Baseline	Couche de base à manches courtes pour hommes	57.1	0
	Whole glossary	*Première couche* à manches courtes pour hommes	**14.3**	**100**
2	Source	Our **helmets** combine lightweight construction and [...] our **EPS** technology		
	Reference	Nos *casques* [...] et de notre technologie *EPS* [...] tout en restant légers.		
	Whole glossary	Nos *casques* combinent une construction légère [...] avec notre technologie de mousse d'absorption EPS [...]	62.5	100
	TM-based filter >40%	Nos *casques* combinent une construction légère [...] avec notre technologie *EPS* [...]	**55**	100
	Manual filter	Nos *casques* combinent une construction légère [...] avec notre technologie *EPS* [...]	**57.5**	100
3	Source	Choose [...] based on boot [...] and desired on-**snow feel**.		
	Reference	Ajustez le niveau [...] de chaussure et du *toucher de neige* recherché.		
	Baseline	Choisissez les modes [...] de la chaussure et des sensations souhaitées sur la neige.	69.6	0
	Whole glossary	Choisissez les modes [...] du boot et du *toucher de neige* souhaité.	**56.5**	**100**
	Rule-based filter	Choisissez les modes [...] de la boot et du *toucher de neige* souhaité.	**52.2**	**100**
	TM-based filter >40%	Choisissez les modes [...] de la chaussure et du *toucher de neige* souhaité.	**43.5**	**100**
4	**Source**	- zur **Installation** von drei TFT-/LCD-/LED-Monitoren mit einer **Bildschirmdiagonale** von 33 bis 69 cm (13" bis 27")		
	Reference	- For the *installation* of 3 TFT/LCD/LED monitors with a *screen diagonal* of 33 to 69 cm (13" to 27")		
	Baseline	- for *installation* of three TFT/LCD/LED monitors with a *screen diagonal* of 33 to 69 cm (13" to 27")	7.1	100
	Whole glossary	- for *Installation* of three TFT/LCD/LED Monitors with a *Screen Diagonal* of 33 to 69 cm (13" to 27")	10.7	100
	TM-based filter >100%	- for the *installation* of three TFT/LCD/LED monitors with a *screen diagonal* of 33 to 69 cm (13" to 27")	**3.6**	**100**
5	Source	Bei Erreichen der max. Lautstärke hören Sie einen **Signalton**.		
	Reference	Once the max. volume is reached, a *signal tone* is heard.		
	Baseline	When the max. volume is reached, you will hear a beep.	37.5	0
	Whole glossary	When the max. Loudness is reached, you will hear a *Signal Tone*.	50.0	**100**
	Manual filter	When the max. volume is reached, you will hear a *Signal Tone*.	43.7	**100**

Table 5: Examples of sentences with their corresponding TER and accuracy scores for both Electrical devices De–En and Sportswear En–Fr using provider1.

without any change in the accuracy. Although the whole glossary and the filtered one might contain terms that are not in the validated glossary used to compute accuracy, and therefore there might be terminology changes that are not captured by the accuracy score, this might also suggest that the use of the terminology feature is introducing some side effects to the sentence translation. This will be further investigated in Sect. 4.3.

For Electrical devices, in Table 4 we observed that TER worsened where accuracy was higher, which was especially true for the whole glossary, and the rule-based and manual filters. Fig. 1 displays that these three glossaries have the highest number of sentences where both TER and accuracy improved, which is the desirable result for these experiments. However, all three outputs show a notable amount of sentences where accuracy improved, but TER either regressed or remained unchanged compared to the baseline. This, together with the fact that the whole glossary and the rule-based filtered one have very similar results, suggests that the latter has removed many of the entries that would not match in the text, e.g. because they contain information in the term field, but some of the terms that should have been removed because of their negative impact on the output were kept (see example 5 in Table 5).

For Electrical devices, the TM-based one removes more terms than the other filters, thus almost 88% of its sentences are the same as in the baseline. The amount of sentences where both

TER and accuracy improve is limited, whereas for 9% of its sentences, TER changes are observed while accuracy stays the same as in the baseline.

As for Sportswear, in Fig. 1 more than half of the sentences for all the outputs show an improvement in TER with either unchanged or improved accuracy (see *Both improved* and *only TER improved* categories), and very few sentences where the accuracy either improved or remained unchanged while TER regressed. This validates how a higher accuracy brings a better TER for the majority of the outputs.

The majority of the sentences in Electrical devices were the same as the baseline, especially with the TM-based filter, which has the best TER (see Table 4). Given that the performance of the baseline is already good, this can be seen as a desirable effect of filtering a glossary. Sportswear, on the other hand, had a poor baseline, especially in terms of accuracy. We thus expected the glossary to impact more sentences, which is what happened. Also, the number of sentences where TER improves is definitely larger than the number of observations where TER regresses, which is particularly true for the TM-based filter 40%.

4.3 Manual analysis

Table 5 displays examples taken from both use cases, along with their accuracy and TER scores. As in the preceding section, we are limiting our scope to sentences produced by Provider 1. For the sake of readability, we are not reporting all candidates for each source, and some sentences have been shortened. The source glossary matches are in bold, and their target (if any) is italicized in the target sentences when correct, and underlined if enforced but incorrect.

Example 1 depicts a scenario of the best result that may be reached using a glossary, i.e. an improvement in both accuracy and overall quality. The baseline did not translate the term accurately, while thanks to the glossary, TER dropped to 14.3% and the accuracy increased to 100%.

Example 2 and 3 demonstrate scenarios in which accuracy does not change while TER changes. In Example 2, the source term *helmet* is translated correctly in all sentences. However, the glossary translation of *EPS* is "mousse d'absorption EPS", and it is only found in the whole glossary. This target term should not have been enforced (see reference). Thanks to the correct decision to remove it from the glossary, both the TM-based filter and the manual filter improved in terms of TER, but the accuracy score did not change since the entry is not in the validated glossary used to compute it. In Example 3, despite the fact that all glossaries correctly introduced the sole entry matching on the source, all candidates are distinct, which shows that the glossary features are introducing side effects.

Examples 4 and 5 are taken from Electrical devices De–En. The former shows a sentence where the rather generic term "installation" was in the whole glossary. This term was filtered out from the TM-based filtered glossary (and also from the rule-based filtered glossary, not appearing here). Although the difference is minor, not having the term in the glossary improves the sentence thanks to the insertion of the article before "installation". This pattern is even more evident in example 5, where another generic term pair ("Lautstärke" translated as "Loudness") is enforced in the whole glossary output, while in this context "volume" would be the correct translation. The term pair including "Loudness" was correctly removed from the manually filtered and the TM-based filtered glossaries. This is one possible explanation for the cases of sentences where TER changes and accuracy does not, as seen in Sect. 4.2.

5 Conclusions and future work

In this paper, we used various approaches to filter pre-existing glossaries and tested their usefulness in improving the terminology accuracy of an MT output without deteriorating the overall quality. The results show that using a filtered glossary may produce a better accuracy with similar or improved overall quality when compared to a baseline where no glossary is used. In several cases a filtered glossary led to a better TER than the whole glossary, which suggests that filtering removes matches from terms that are harmful for the overall quality. On the other hand, results show that using a whole glossary can be beneficial. The whole glossary usually outperformed the filtered ones in terms of accuracy – which is expected given the larger size of the former – and, especially in the case of Sportswear En–Fr, the TER improvements brought by filtering were rather small. In general, filtering – and in particular the TM-based automatic filter – helped find an acceptable balance between a higher accuracy and a good overall quality. Given the current experimental stage of the filtering tool, such results can be seen as promising. However, improvements to the filtering method and new tests are needed to check if filtering can bring larger quality improvements over the whole glossary, thus making the filtering effort worthwhile.

The results in both use cases suggest that aiming at the highest possible accuracy may not always be the best choice in terms of quality. There appears to be an accuracy cut-off beyond which overall quality suffers. In the case of Electrical devices, this could be due to the fact that the terminology is quite general – indeed, the baseline is already handling the majority of the terms correctly. The analysis in Sect. 4.2 revealed that the TM-based filtering method with the most restrictive threshold introduced only minor changes with respect to the baseline output, reducing the number of sentences where TER was regressing. This behavior may be preferable to using a larger glossary, which can negatively impact more sentences, especially when the baseline is performing well in terms of accuracy and overall quality.

The baseline in Sportswear En–Fr is struggling with terminology accuracy, indicating that terminology is highly domain/customer-specific. In terms of TER, two automatic filters outperform the whole glossary, whereas the manually filtered glossary achieves the highest accuracy, closely followed by the whole and rule-based glossaries. Applying a strict filter does not improve the quality of these contents. When compared to the baseline, we discovered that each glossary affects at least 70% of the sentences (Sect. 4.2). However, we still see an accuracy cut-off around 70% (see Table 4), above which TER begins to deteriorate. This may imply that, while including as many terms as possible may be desirable, applying a filter to the glossary can help removing some that are detrimental to the overall quality.

An interesting conclusion we can draw from our experiments is that a glossary filtered by a linguist according to task-specific guidelines does not necessarily bring relevant improvements over an automatically filtered glossary. In particular, the TM-based filter always outperforms the manual one in terms of TER score. The rule-based filter outperforms the manually filtered glossary for Sportswear En–Fr in terms of TER, and achieves a slightly lower accuracy score. Given the high costs of manually filtering a glossary, this can be considered a relevant outcome, especially for LSPs. In some cases, even for a linguist expert of the content type, it can be difficult to distinguish a generic term from a specific one, which is one of the key actions to take when filtering a glossary for MT.

Although results suggest that using a TM to identify terms that are not highly specific to one domain can be effective, we plan to test more accurate solutions to this problem, such as the use of Inverse Document Frequency (IDF) (Jones, 1972) or word-alignment. In Bergmanis and Pinnis (2021a), both methods were tested for glossary filtering. We anticipate that an improved ability to filter out generic terms will be especially helpful in use cases such as the one of Electrical devices De—En.

The rule-based filter, which requires no bilingual data other than the glossary, has one of the highest accuracy and one of the best TER scores in Sportswear En—Fr. This result is especially relevant in cases where a glossary must be filtered but bilingual data are either not available or their quantity is limited.

Examples shown in Sect. 4.3 suggested that there can be several reasons for quality improvements or regressions when terminology is added to the output. Sometimes the output changes even if no term was matched in the sentence, which is probably due to the specific implementation of the glossary feature. To gain a better understanding of this, we plan to carry out in-depth manual analyses of the outputs produced by the baseline, by the whole glossary and by the filtered ones.

The present contribution focused on term filtering. However, the ability to edit terms that can be improved may yield better results. In the future, we will concentrate on this, beginning with cases where terms can be easily improved using automatic editing rules. Editing terms without the assistance of a linguist can be a difficult task at times. We therefore intend to conduct experiments in which the results of the automatic filters are provided to a linguist as an aid to help them perform their task. We anticipate that this will also help linguists in their decision making process, e.g. to determine when terms are generic. Being able to see that a term is translated inconsistently in the TM, for example, can lead to a better decision as to label the term as generic/detrimental or not.

Acknowledgements

The authors would like to thank the whole Acolad AI R&D team for their valuable feedback, and three anonymous reviewers for insightful comments on the first draft of this paper.

References

Ailem, Melissa, Jinghsu Liu, and Raheel Qader. 2021. Encouraging Neural Machine Translation to Satisfy Terminology Constraints. *arXiv:2106.03730 [cs]*, June. arXiv: 2106.03730.

Alam, Md Mahfuz ibn, Antonios Anastasopoulos, Laurent Besacier, James Cross, Matthias Gallé, Philipp Koehn, and Vassilina Nikoulina. 2021a. On the Evaluation of Machine Translation for Terminology Consistency. *arXiv:2106.11891 [cs]*, June. arXiv: 2106.11891.

Alam, Md Mahfuz Ibn, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Kweonwoo Jung, Philipp Koehn, and Vassilina Nikoulina. 2021b. Findings of the WMT shared task on machine translation using terminologies. In *Proceedings of the Sixth Conference on Machine Translation*, pages 652–663, Online, November. Association for Computational Linguistics.

Bergmanis, Toms and Mārcis Pinnis. 2021a. Dynamic Terminology Integration for COVID-19 and other Emerging Domains. *arXiv:2109.04708 [cs]*, September. arXiv: 2109.04708.

Bergmanis, Toms and Mārcis Pinnis. 2021b. Facilitating Terminology Translation with Target Lemma Annotations. *arXiv:2101.10035 [cs]*, January. arXiv: 2101.10035.

Bergmanis, Toms, Mārcis Pinnis, and Paula Reichenberg. 2021. From research to production: Fine-grained analysis of terminology integration. In *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*, pages 54–77, Virtual, August. Association for Machine Translation in the Americas.

Chatterjee, Rajen, Matteo Negri, Marco Turchi, Marcello Federico, Lucia Specia, and Frédéric Blain. 2017. Guiding neural machine translation decoding with external knowledge. In *Proceedings of the Second Conference on Machine Translation*, pages 157–168, Copenhagen, Denmark, September. Association for Computational Linguistics.

Dinu, Georgiana, 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, July. Association for Computational Linguistics.

Dougal, Duane K. 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, May. European Language Resources Association.

Exel, Miriam, 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, November. European Association for Machine Translation.

Guerrero, Lucía. 2020. NMT plus a bilingual glossary: does this really improve terminology accuracy and consistency? Slides presented at the Translating and the Computer conference (TC42 online) organised by AsLing (International Association for Advancement in Language Technology), `https://bit.ly/3vTSjCh`, November. Accessed: 2022-03-08.

Hasler, Eva, 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, June. Association for Computational Linguistics.

Hokamp, Chris 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, July. Association for Computational Linguistics.

Jones, Karen Spärck. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1).

Riediger, Hellmut. 2018. *Cos'è la terminologia e come si fa un glossario*. `http://www.term-minator.it/corso/doc/mod3_termino_glossa.pdf`.

Scansani, Randy and Loïc Dugast. 2021. Glossary functionality in commercial machine translation: does it help? a first step to identify best practices for a language service provider. In *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*, pages 78–88, Virtual, August. Association for Machine Translation in the Americas.

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