Glossary functionality in commercial machine translation: does it help? A first step to identify best practices for a language service provider

Randy Scansani rscansani@acolad.com
Loïc Dugast ldugast@acolad.com
NLP team, Acolad

Abstract

Recently, a number of commercial Machine Translation (MT) providers have started to offer glossary features allowing users to enforce terminology into the output of a generic model. However, to the best of our knowledge it is not clear how such features would impact terminology accuracy and the overall quality of the output. The present contribution aims at providing a first insight into the performance of the glossary-enhanced generic models offered by four providers. Our tests involve two different domains and language pairs, i.e. Sportswear En–Fr and Industrial Equipment De–En. The output of each generic model and of the glossary-enhanced one will be evaluated relying on Translation Error Rate (TER) to take into account the overall output quality and on accuracy to assess the compliance with the glossary. This is followed by a manual evaluation. The present contribution mainly focuses on understanding how these glossary features can be fruitfully exploited by language service providers (LSPs), especially in a scenario in which a customer glossary is already available and is added to the generic model as is.

1 Introduction

Correctly translating terminology is one of the main challenges in translation, and this is also true for Machine Translation (MT). A first approach to achieve this goal lies in data preparation and possibly appropriate training algorithms. However, there are cases in which data is not available or training a model is not an option.

A number of research works have explored ways to combine a bilingual glossary with an MT model at run-time, enforcing specific terminology in the output, e.g. Arthur et al. (2016); Chatterjee et al. (2017); Farajian et al. (2018); Hasler et al. (2018); Dinu et al. (2019); Exel et al. (2020); Bergmanis and Pinnis (2021). The proposed approaches range from simple post-translation replacement, to constrained decoding, down to methods that allow for soft constraints and are able to generate inflected forms of glossary terms – Dinu et al. (2019) improved by Bergmanis and Pinnis (2021). Such recent breakthroughs might not have made it yet to commercial implementation.

Nevertheless, a number of commercial MT providers have started to offer features allowing users to enhance a generic MT model by leveraging a bilingual glossary.[1] While language

[1]Some examples of MT providers offering a glossary feature: DeepL (`https://bit.ly/2UbHDyh`), Google Translate (`https://bit.ly/3rcUqwv`), Microsoft (`https://bit.ly/2U5os9v`), Amazon Translate (`https://amzn.to/3hC7WGO`), Systran (Michon et al., 2020).

service providers (LSPs) have to rely on those solutions, "the commercial providers usually leave us in the dark about the technology that is used for the implementation of that feature" (Exel et al., 2020).

Users may expect the addition of a domain-specific glossary to a generic MT model to bring improvements both in terminology translation and, as a result, in the overall output quality. The present contribution aims at understanding if the glossary features offered by some of the main MT providers are meeting such expectations. More specifically, being a relevant scenario for LSPs, we aim at understanding if a glossary extracted from a customer termbase can be leveraged as is, i.e. all experiments will be carried out using the glossaries without any preliminary cleaning up. We will further refer to this use case as *naive use of glossary*. Four different MT providers will be tested in the sportswear (En–Fr) and in the industrial equipment (De–En) domains, comparing their performance when the glossary feature is switched on and when it is not.

More in detail, the impact of the glossary feature on terminology translation will be assessed by checking the extent to which the MT output complies with the glossary entries. A first evaluation will follow strict parameters, i.e. glossary term matching is case-sensitive and happens on a token level. We will refer to this evaluation as *exact match*. In a second evaluation (henceforth *loose match*), we aim at finding any terminology improvement by matching terms on a lemma level and without considering differences in casing. The effect of the glossary on the overall output quality will be measured with Translation Error Rate (TER) (Snover et al., 2006). Based on term matching and on TER, we will then categorize each sentence based on the terminological and/or qualitative improvements (if any). To conclude, a manual evaluation will provide a more detailed overview on the glossary impact on the sentence.

The aim of the contribution is to start addressing the needs for best practices across the translation industry for the use of glossaries to improve MT output. Given the availability of glossary features, how can we leverage pre-existing glossaries?

The remainder of the present contribution is structured as follows. In Sect. 2, a description of the experimental setup will be provided, including descriptions of the MT providers, the data sets, the metrics and the evaluation methods. The following Section (Sect. 3) will present the results obtained with the *naive use of glossary* approach. First we will focus on each provider's behavior on the whole data sets (Sect. 3.1), then a sentence level analysis is carried out (Sect. 3.2), and finally we will present the results of a manual annotation (3.3). This is followed by a discussion of the results obtained (Sect. 4).

2 Experimental setup

2.1 Machine translation providers

In the present contribution, 4 providers were tested, comparing the performance of their generic model against the same model enhanced with the glossary functionality. We are not providing the name of the 4 engines since this paper does not claim to present an exhaustive benchmarking, but rather aims at investigating how to make the best out of such glossary functionalities in a scenario relevant to the language industry.

All MT providers disclose only a limited number of details on how the terms are matched and enforced into the output. The glossary feature of Provider 2 and 4 is described as a simple replacement of the target term(s) generated by the model with the one(s) included in the glossary, whenever a glossary item is matched in the source text. Provider 2 further specifies that the rest of the sentence is not adjusted after the term enforcement. To the best of our knowledge, Provider 1 and 3 have not published any technical specifications on their glossary feature.

Regarding the recommendations available, Provider 1, 2 and 4 indicate that the glossary feature is especially useful to enforce the preferred translation for product names and/or non-

context dependent source terms for which we want to enforce a unique domain-specific translation. Provider 2 and 4 further indicate that the glossary functionality is case-sensitive, so the glossary term must match the casing used in the text.

Providers 1, 3 and 4 allow to specify a glossary at translation time, which will be enforced during translation. Provider 2 offers two different options. With the first one – henceforth referred to as Provider 2-preprocess – the source text can be preprocessed to tag glossary terms so that they can be identified by the model at translation time. With the second one – henceforth referred to as Provider 2-pretrained – a training is launched using a glossary as unique training data set. Even though there is a training step involved, the provider specifies that this option simply replaces the terms in the output with those included in the glossary.

In order to have better insight into how the different providers match terms in the source and enforce their translation in the target, we run some preliminary tests. The information retrieved from the tests and from the specifications mentioned above are summed up in Table 1. It is worth noting that for Provider 2-preprocess, its case-sensitivity on the source side does not depend on the provider specifications, but rather on the preprocessing method implemented by the user. In our case, the preprocessing procedure that tags source terms in the text is case-insensitive.

Provider	**Source matching**		**Target insertion**
	Case-sensitive	**Matches lemmas**	**Sent. adjusted**
Prov. 1	✓	✗	✗
Prov. 2-pretr.	✓	✗	✗
Prov. 2-preproc.	✗	✗	✗
Prov. 3	✗	✓	✓
Prov. 4	✓	✗	✗

Table 1: The *Source matching* columns describe how the matching happens in the source text for each provider. The *Target insertion* column specifies which providers adjust the target sentence after enforcing a glossary term.

2.2 Data sets

Two data sets in two different language pairs and domains were extracted for this task, i.e. De–En Industrial Equipment and En–Fr Sportswear. This allows to test the usefulness of the glossary features for two different types of contents. Also, we are interested in the possible differences between one language pair where the source language has more inflections than the target one (De–En), and a language pair where more inflections occurs on the target side (En–Fr).

After extracting a test set from the bilingual corpora of each customer, we select subsets by keeping only sentence pairs containing at least one source-target match from the glossary. A description of the matching method is provided in Sect. 2.3. The test set and glossary size are shown in Table 2. In this *naive use of glossary* approach (see Sect. 1) we are not preprocessing the two glossaries. However, one of the providers used does not allow multiple entries with the same source term. For this reason, we chose to randomly pick one of the target terms and discard the other ones. 51 entries were removed from the En–Fr glossary, while the De–En one did not contain any source duplicates.

2.3 Metrics

The different analyses carried out in this paper are focused on assessing the extent to which the outputs comply with the glossary terminology and on evaluating the overall output quality.

Domain	Source	Target	Term pairs	Sent. pairs
Industrial equipment	DE	EN	345	1063
Sportswear	EN	FR	1708	1673

Table 2: Domain, source and target language, number of term pairs and sentence pairs available for each data set used.

For the latter, we use (case-sensitive) TER (Snover et al., 2006), while the former assessment is performed by a specific script described in Algorithm 1.

Optional: Lemmatize terms in glossary and sentences in test set ;
Optional: Lowercase terms in glossary and sentences in test set ;
Find all occurrences of source terms;
Disambiguate overlapping source terms (choose longest entries first);
Count matches in the target language sentences;
Result: Match accuracy

Algorithm 1: Compute term matches within candidate translation

2.4 Automatic analyses method

In the first analysis, whose results are described in Sect. 3.1, the number of source and target terms matched by the algorithm is used to compute accuracy as in Alam et al. (2021), i.e. the proportion between the number of source terms whose target is matched in the target text and the total number of matched source terms.

While the first analysis provides insight into the performance of each provider on the whole data set, it does not allow for a more granular understanding of the glossary impact on a sentence level. To this aim, we perform a second analysis where we compare the generic output of each provider to the glossary-enhanced one on a sentence level. Each sentence is assigned to one of the six categories below, according to the accuracy and TER changes observed after the addition of the glossary. These six categories are similar to those suggested in Alam et al. (2021) for the classification of MT systems based on their ability to correctly handle terminology.

	TER (↓)	**Acc. (↑)**
Accuracy or both regressed	$\uparrow$ or $=$	$\downarrow$
TER only regressed	$\uparrow$	$=$
Unchanged	$=$	$=$
TER only improved	$\downarrow$	$=$ or $\downarrow$
Accuracy only improved	$=$ or $\uparrow$	$\uparrow$
Both improved	$\downarrow$	$\uparrow$

Table 3: Description of the six categories used in the sentence-level analysis (results in Sect. 3.2). Any change in the TER or accuracy values is measured comparing the translation of each source sentence by the generic model and by the glossary-enhanced one.

2.5 Manual evaluation method

In order to better assess the effect of the glossary feature, we look into the target sentences to spot any difference between the output of the glossary-enhanced model and that of the generic model. In particular, we want to understand if terminology is inserted in the correct context, and how the rest of the sentence changes. For each category in Table 3, we pick a random set of 10 segments to be manually annotated by one annotator for each language pair. Sentences

belonging to the *Unchanged* category are not annotated.

Differences between the two sentences in each pair are annotated with the following labels, distinguishing between regressions and improvements: `Casing`, `Inflection`, `Word order`, Part-Of-Speech (`POS`), `Terminology`, `Lexical choice`, `Other`. Please note that `Terminology` refers to changes impacting a matched source term and its translation, whereas `Lexical choice` includes any other lexical change.

3 Experimental results

3.1 Accuracy and TER on the whole data sets

Provider	De–En (%)			En–Fr (%)		
	Exact match Acc. ↑	Loose match Acc. ↑	TER ↓	Exact match Acc. ↑	Loose match Acc. ↑	TER ↓
Prov. 1	63.7	85.1	31.6	42.1	45.1	61.3
Prov. 1 + gloss.	99.6	95.8	33.0	95.2	77.7	60.5 †
Prov. 2	57.9	80.9	33.2	33.9	38.2	65.6
Prov. 2-pretr.	99.9*	95.5	34.4	98.6*	78.4	65.4 †
Prov. 2-preproc.	99.9*	97.0	34.1	95.2	87.8*	64.6 †
Prov. 3	45.5	68.9	32.3	43.2	46.0	61.0
Prov. 3 + gloss.	78.1	98.4*	29.9 †	78.6	79.5	59.2 †
Prov. 4	54.7	77.3	34.3	43.0	46.6	63.0
Prov. 4 + gloss.	88.7	93.4	33.9 †	90.9	75.1	61.3 †

Table 4: Accuracy and TER results for each provider with and without glossaries, and for each of the use cases (De–En industrial equipment, En–Fr Sportswear). TER is provided only once since the test set for the two evaluations is the same. † identifies a TER decrease when the glossary is added. * identifies the providers with the best accuracy.

Exact match In this evaluation, terms are matched on a token level (no lemmatization) and only when the casing in the output is the same as the one in the glossary. Results in Table 4 (*Exact match* and *TER* columns) show that Provider 1 and Provider 2 achieve the highest accuracy scores (99.9%) for both use cases (De–En Industrial Equipments and En–Fr Sportswear). However, the glossary impact on the overall quality is not on par. For De–En the use of the glossary decreases TER for Provider 3 and 4 only. For En–Fr TER always decreases when a glossary is added to the generic model, although some of these drops are rather limited, ranging from -0.18% (Provider 2-pretrained) to -1% (Provider 2-preprocess). Exact match accuracy for Provider 3 and 4 enhanced with glossary is lower than Provider 1 and 2 with glossary. This is expected since Provider 3 glossary feature is able to generate a different target inflection, which is not recognized in the exact matching. The quality increase is however larger. For example, we observe a -2.4% TER when a glossary is added to Provider 3 for De–En, and a 1.8% TER drop for the same provider on En–Fr.

Loose match In this evaluation, terms are matched on a lemma level and regardless of their casing. With respect to the first evaluation, this brings a higher accuracy for the generic models without glossary (see Table 4), which means that the generic models are often using the correct lemma. Provider 3 achieves the best accuracy for De–En (98.4%) and the 2nd best accuracy for En–Fr (79.5%), narrowing the gap with the best-performing model (Provider 2-preprocess, 87.8%) wrt the exact match results. Provider 2-preprocess accuracy drop from the exact match

to the loose match evaluation is less evident than the accuracy drop of Provider 2-pretrained (which is case-sensitive) for both De–En and En–Fr. For En–Fr we observe a large accuracy drop wrt the previous evaluation for all Providers except Provider 3, due to its ability to match different inflections of a source term. Provider 1, 2 and 4 match source terms on a token level (see Table 1). To conclude, in this evaluation we see that Provider 3 has the best TER scores in both language pairs. As seen above, while TER always decreases when a glossary is added to the En–Fr models, for De–En the same happens only for Provider 3 and 4.

3.2 Sentence-level analysis

In this analysis we are comparing, for each provider, the output of the generic model to the output of the glossary-enhanced one. Sentences are assigned to one of the categories described in 2.4.

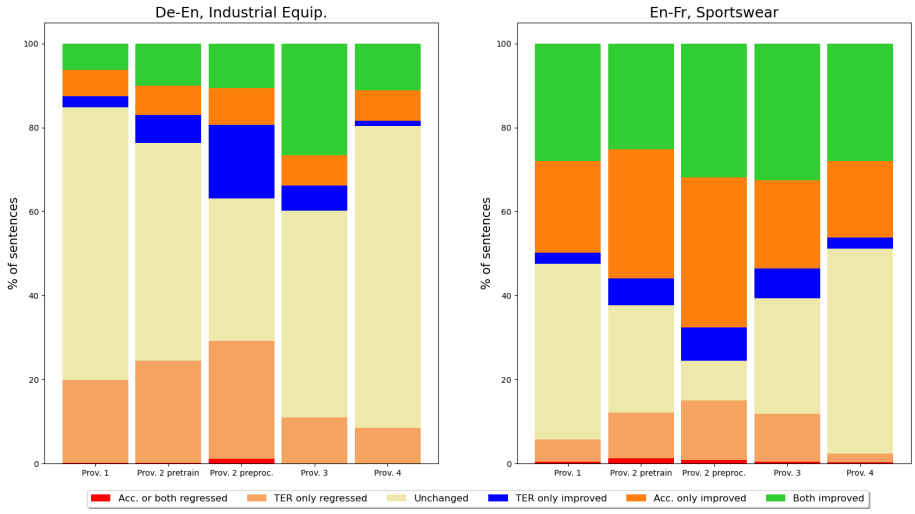

Figure 1: For each use case (De–En Industrial Equipment and En–Fr Sportswear) we report on the percentage of sentences produced by each provider that were assigned to one of the six categories described in Table 3. The six categories refer to the comparison between the generic model of a provider and its glossary-enhanced version.

As can be seen in Fig. 1, there are large differences between the two language pairs, the most evident being perhaps the higher quantity of Unchanged sentences for De–En. This could be motivated by the differences in the size of the two glossaries (see Table 2), but also by the differences between the two language pairs. Given the morphological complexity of German, matching the correct form of the term in the source text is more difficult than for English.

The percentage of sentences where only accuracy improved is higher for En–Fr than for De–En. This seems to suggest that using the glossary introduces more side effects if the target language has more inflections and concordances, thus TER does not improve.

Comparing the providers, Provider 2-preprocess seems to have the highest number of side effects, since the percentage of sentences where TER only improves or TER only regresses is the highest in both language pairs. Provider 3 (De–En) shows the highest percentage of sentences where both TER and accuracy improved, and the highest amount of sentences where the use of the glossary was beneficial (i.e. sentences assigned to *TER only improved*, *Accuracy only improved* or *Both improved*). For En–Fr Provider 2-preprocess shows the highest number

of sentences where the use of a glossary had a positive impact, although the percentage of sentences belonging to *Both improved* is the same as that of Provider 3.

Provider 1 and 4 show similar performances when it comes to the sentences where the glossary-enhanced model is beneficial to either accuracy, or TER or both of them in both use cases. However, Provider 4 is the Provider with the lowest portion of sentences where TER only regresses. To conclude, the differences between Provider 2-pretrained and Provider 2-preprocess show that the latter approach is more effective.

3.3 Manual annotation

We chose to limit the annotation to Provider 2-preprocess and Provider 3, due to the good performance shown by both in the previous analyses (see Sect. 3.1 and 3.2), while their specifications differ. Since Provider 3 is able to handle morphology inflections, its impact on the output sentence might differ from that of the other providers. Also, we wanted to have a better understanding of Provider 2-preprocess performance given the apparently high number of unexpected behaviors of such provider, i.e. the high number of sentences where TER only regressed or TER only improved seen in Fig. 1.

Looking at Table 5, one of the most evident results is that most of the improvements for both providers in both language pairs are due to terminology. This confirms the results seen in Sect. 3.1 and 3.2, i.e. the glossary features does increase the amount of correct terms in the output.

As expected, we see a high number of both positive and negative side effects for both providers. For Provider 2-preprocess we see many side effects in different categories, especially in En–Fr, many of which are negative (see for example the `Inflection`, `POS` and `Word order` columns). Example A in Table 6 shows a casing issue and a wrong concordance. The term "noyeau" was specified in the glossary as the translation for *core* (both lowercased). It has to be reminded that `casing` issues are also due to our choice to tag terms in the source text regardless of their casings (see Sect. 2.1). This has increased the number of glossary matches, but it might have increased the number of casing issues in the target as well.

However, some `casing` issues are not due to the glossary. In En–Fr, Example B (Table 6) sees the MT lowercasing the whole sentence. The glossary contains a single entry for *outdoor*, which is lowercased and where the source English term is copied to the target side.

Some glossary entries were actually not valid terms, and their enforcement in the output might have harmed the translation quality/correctness in some cases. Indeed, we see a number of `terminology` regressions for both providers. In example C, Provider 3 produced a wrong translation because of a glossary entry that included a preposition, i.e. "NEXT" as the translation of "Vor".

Differences in terms of lexicon between the generic output and the glossary-enhanced one were annotated as `lexical choice` improvements or regressions, provided that such differences were not caused by the use of the glossary. As can be seen in Table 5, this class has many examples across all sentence categories and providers. In Example D (Table 6), although small, the translation of "Betriebsart" as *Operating mode* can be considered an improvement since the target term matches the source one exactly, while *mode* is a correct translation but not as accurate. These words were not included in the glossary. In Example E the translation for the German conjunction "weswegen" is missing, so the meaning of the sentence is not correctly conveyed.

Differences between the two language pairs can also be observed. For example, for De–En we see a higher number of `casing` regressions and improvements, probably due to the mismatch between the German and the English casing (see Example B, discussed above). Even more evident are the differences between the amount of `inflection` issues (and some im-

Provider, Lang. pair	Sent. Category	Casing	Infl.	Word ord.	POS	Term.	Lex.	Oth.
2 preproc., De–En	Acc. or both regressed	+++ ---				--	+++ --	
	TER only regressed	+ --		+ -	-		+ ---	
	TER only improved	+ -		+++			+++ ---	+
	Acc. Only improved	+ ---		 -		+++++ -	+ --	
	Both improved	++ -		+ -		+++++ -	+++ -	
3, De–En	Acc. or both regressed							
	TER only regressed	++ -	+	+ --	+	 -	 ---	
	TER only improved	++		+			+++ ---	
	Acc. Only improved	 -				++++ --	++ ---	
	Both improved			+		+++++	 -	

Provider, Lang. pair	Sent. Category	Casing	Infl.	Word ord.	POS	Term.	Lex.	Oth.
2 preproc., En–Fr	Acc. or both regressed	+ --	 ---		 --	+	++ -	
	TER only regressed	 --	 --		 --	+	++ -	
	TER only improved	+ -	+ -	 --	 -	++	+++ -	
	Acc. Only improved	 -	 --	 --		+++++ --	+ -	
	Both improved		 --	+ --		+++++ --	++ --	
3, En–Fr	Acc. or both regressed	 -				 ----	 -	
	TER only regressed	 ---					+ --	
	TER only improved	+ --		+ -		+	+ + + --	
	Acc. Only improved		+		-	++++ -		
	Both improved		 -		 -	+++++ -	+	

Table 5: Results of the manual annotation on De–En (above) and En–Fr sentences produced by Provider 2-preprocess and Provider 3. The amount of errors in each error class (columns) was normalized over the number of sentences in that category (row). The higher the number of + or -, the higher the number of, respectively, improvements or regressions.

provements) for En–Fr *vs.* De–En, which is obviously due to the higher number of inflections and concordances in the French language.

At the same time, for En–Fr we see that the number of `inflection` issues is reduced for Provider 3. The same can be observed, e.g., for the `word order` class. For Provider 2-preprocess (En–Fr) `word order` regressions were seen in three sentence categories (TER only improved, Accuracy only improved and Both improved). For Provider 3 we see `word order` regressions in one category only (TER only improved).

Looking at differences between sentence categories, when there are regressions we can see a different number of causes, e.g. `lexical choice` regressions, `casing` regressions or `inflection` regressions (especially for En–Fr). On the other hand, the three categories where the glossary-enhanced output is better (TER only improved, Accuracy only improved or Both Improved) are highly influenced by `terminology` improvements, as can be seen by the high number of + symbols in the `terminology` column. An example of sentence where both accuracy and TER improved is example F in Table 6. Here, the use of a glossary term caused the output to be more similar to the reference text, which caused a TER decrease.

Ex.	Prov.	Gloss.	Sentence
A	source		(...) ABOVE *THE CORE*
	2-preproc.	✗	(...) AU-DESSUS *DU NOYAU*
		✓	(...) AU-DESSUS *DE LA noyau*
B	source		OUTDOOR GEAR LAB - TOP PICK
	3	✗	OUTDOOR GEAR LAB - TOP PICK
		✓	outdoor gear lab - premier choix
C	source		*Vor* der erstmaligen Wartung (...)
	3	✗	*Before* the unit is serviced for the first time (...)
		✓	*NEXT*, when the device is serviced for the first time (...)
D	source		*Betriebsart Timer* nicht möglich (...)
	2-preproc.	✗	*Timer mode* not possible (...)
		✓	*Operating mode Timer* not possible (...)
E	source		(...), *weswegen* in den letzten Jahren viele Projekte zur Wassergewinnung geplant wurden, (...)
	2-preproc.	✗	(...), many water extraction projects have been planned, (...)
		✓	(...), *which is why* many water extraction projects have been planned, (...)
F	source		Die Einstellungen am *Gerät* sind (...)
	3	✗	The settings on the *unit* are (...)
		✓	The settings on the *device* are (...)

Table 6: Examples of sentences from the two data sets (En–Fr and De–En). Italics is used to highlight the parts of the sentences that are discussed in Sect. 3.3.

4 Conclusion and future work

The experiments described in the previous sections illustrate how the naive use of a glossary may not always provide the expected outcome, i.e. a better terminological compliance together with an overall improved output quality. Results depend on the implementation of the glossary feature by the MT provider (how entries are matched and enforced on the target side), on the language pair and on the glossary itself.

Regarding the differences between providers, those that are able to handle morphology

(Provider 3) have shown to produce more sentences where terminology improvements result in a better overall quality. Most implementations seem to induce a number of undesirable side-effects on casing, morphology, word order. Moreover, some limitations remain for all providers tested. For example, none of them (including Provider 3) is able to match glossary source terms when these occur in a compound term (e.g. matching the German term *Batterie* if the source text contains *Batterietyp*). This would impact all agglutinative languages.

Besides the specifications of the glossary features, we saw that some glossary entries brought a lower translation quality, which raises questions about the quality of the glossary itself (see example C in Table 6). For instance, a glossary might have been created by the customer without the support of any terminologist – e.g. to update and/or validate the entries – and then provided to the LSP. As a result, the termbase might contain more target options for the same source term, or it might include entries that are not relevant (e.g. function words), or entries not domain-specific, whose POS is ambiguous or whose translation is highly context-dependent, even within a well-defined domain.

Starting from the assumption that a customer glossary as is does not comply with some of the specifications set by the providers (see Sect. 2.1), and focusing on a scenario in which we already chose which provider to use, how could we turn a preexisting termbase into an MT-compatible glossary? A manual revision of the whole glossary may be time-consuming and might not solve all issues. As mentioned by Bergmanis and Pinnis (2021), we cannot expect the user to provide for each entry all casing forms, and even less so all inflected forms. Automatic POS tagging could help identifying non-inflective entries, but will be prone to errors.

On the one hand, in order to adapt to the currently available technology, LSPs may have to define best practices. In future work, we intend to run similar tests with subsets of the client glossaries containing only entries that are compliant with the MT providers specifications. Such tests would involve the assessment of different procedures and tools to clean up glossaries. Besides being able to discard entries that are not relevant, a further step would be that of enhancing the glossary by identifying new terms that, if added to the entries, would bring further benefits to the output quality.

On the other hand, the results of the recent research endeavours in the filed of terminology and MT are expected to build momentum for new implementations in commercial solutions, which should narrow the gap between what is currently offered by MT providers and what LSPs are expecting.

References

Alam, M. M. I., Anastasopoulos, A., Besacier, L., Cross, J., Gallé, M., Koehn, P., and Nikoulina, V. (2021). On the evaluation of machine translation for terminology consistency. *CoRR*, abs/2106.11891.

Arthur, P., Neubig, G., and Nakamura, S. (2016). Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567, Austin, Texas. Association for Computational Linguistics.

Bergmanis, T. and Pinnis, M. (2021). Facilitating terminology translation with target lemma annotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3105–3111, Online. Association for Computational Linguistics.

Chatterjee, R., Negri, M., Turchi, M., Federico, M., Specia, L., and Blain, F. (2017). Guiding neural machine translation decoding with external knowledge. In *Proceedings of the Second Conference on Machine Translation*, pages 157–168, Copenhagen, Denmark. Association for Computational Linguistics.

Dinu, G., Mathur, P., Federico, M., and Al-Onaizan, Y. (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.

Exel, M., Buschbeck, B., Brandt, L., and Doneva, S. (2020). Terminology-constrained neural machine translation at SAP. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 271–280, Lisboa, Portugal. European Association for Machine Translation.

Farajian, M. A., Bertoldi, N., Negri, M., Turchi, M., and Federico, M. (2018). Evaluation of terminology translation in instance-based neural MT adaptation.

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

Michon, E., Crego, J., and Senellart, J. (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.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (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.