What's on an annotator's mind? Analysis of error typologies to highlight machine translation quality assessment issue

Emmanuelle Esperança-Rodier

What's on an annotator's mind?
Analysis of error typologies to highlight machine translation quality assessment issue

Emmanuelle Esperança-Rodier
Univ. Grenoble Alpes, CNRS, Grenoble INP[1], LIG, 38000 Grenoble, France
Emmanuelle.Esperanca-Rodier@univ-grenoble-alpes.fr

Abstract

1. Introduction

At the era of Artificial Intelligence, billions of words are needed to train, fine-tune and test Neural Machine Translation systems (Koehn and Knowles, 2017; Linzen, 2020). Quality assessment of the models therefore requires the creation of a lot of annotated corpora. Several studies questioned the quality of those corpora, in terms of reliability and reproducibility in Bregeon et al. (2019), comparability in Echart et al. (2012), who claimed that "the quality of statistic measurements on corpora is strongly related to […] corpus quality"; as well as in Mathet et al. (2012) who wrote that "the quality of manual annotations has a direct impact on the applications using them".
This preliminary work addresses the quality of the annotated corpora. Analysing how one single annotator behaves while annotating the same documents with two different typologies, we show that the choice of a typology impacts the results. Popovics (2018) raised issues on the identification of patterns causing translation errors and we show in this paper that patterns are not identified in the same way depending on the typology used. After presenting the data, we will analyse them under quantitative and qualitative measures.

2. Methodology

2.1 Corpus

Focusing on the quality, we selected a corpus of 111 segments made of 791 source words (English) and 2 151 target words (French).
Firstly, the 9 first segments of a patent document were translated using the WIPO neural and statistical Machine Translation (MT) systems, ending up with 18 translations. Secondly, the 28 first segments of an environmental regulation-like document were translated using the statistical and neural MT systems from the European Commission, respectively MT@EC and eTranslation. From the 28 English source segments, we obtained 56 French translations.

2.2 Annotation

We then, asked an undergraduate student studying Literature and Modern Languages (University of Exeter), a native English speaker and certified C2 in French, to annotate thanks to ACCOLÉ (Esperança-Rodier et al., 2019) the corpus using dVilar's typology (Vilar et al., 2006) and DQF-MQM one (Lommel et al., 2018).
We obtained 137 error annotations using Vilar's typology and 122 using the DQF-MQM one (See Annex 1).

[1] Institute of Engineering Univ. Grenoble Alpes

3. Analysis

3.1 Quantitative Analysis

3.1.1 Global Quantitative Analysis

We have a 16% increase (See Annex 1) in the total number of annotations under Vilar's typology on the statistical translation, reaching 18% for the neural translation when annotating the patent document, compared to a less remarkable increase of respectively 8,5% and 2.8% for the environmental document.

Comparing the two MT systems, more errors are annotated for the statistical systems with a very noticeable gap for the patent document (roughly 120% additional errors for both typologies), than for the neural systems ranging from 46% to 64% additional errors, depending on the annotation typologies.

Looking at the error rank depending on the typologies (See Annex 2), the two first equivalent error types are, for DQF-MQM, Accuracy> Mistranslation (38%) and Addition (16%), and for Vilar's, Incorrect Words> Sense> Wrong Lexical Choice (40%) and Extra Words (17%), with almost the same percentage of use. This complies with Klubicka et al. (2017)'s research in which Mistranslation was the most common of the annotated error types.

Then, the 3rd error for DQF-MQM, Fluency> Grammar> Word Order (9%), is equivalent to two error types in Vilar's, the word level and the phrase level, ranking respectively 4th (7%) and 7th (2%). We already deduce that if weights were associated to the error type for automatic metrics, the comparison of results would be biased as a same error represents one single error type in one typology, and two ones in the other. Hence, to be able to compare the metrics, the typologies have to be aligned to understand their differences when interpreting the metrics.
The 4th ranked error type Accuracy> Omission (7%) behaves identically, as the two equivalent error types in Vilar's, Missing Words> Filler words (4%) 6th rank and Content Words ranked 8th (1%), do not correspond neither in ranks nor percentage. We thus wonder how an annotator behaviour is modified by the possibility of annotating an error upon one or two error types.
This is also illustrated by Vilar's 3rd rank error, Incorrect Words> Incorrect Form (15%) which has two equivalent types in DQF-MQM, Fluency> Grammar> Word form- Part of speech 5th rank (7%) and Agreement 6th rank (6%).
Due to the selection of only Fluency and Adequacy error types for DQF-MQM, Vilar's error type Incorrect Words> Style has no equivalent. This comforts the idea that the choice of the error types before annotating is of the most importance.

3.1.2 Detailed Quantitative Analysis

Looking at the error types ranking for DQF-MQM according to the percentage of errors annotated on the four translations (See Annex 3), there are 62% of additional annotations for the statistical translations compared to the neural ones. Also 42% additional errors were annotated on Robert compared to Jude Law.
We observe the same trend on Vilar's annotations (See Annex 4), with a slight increase of the percentages as more errors were annotated under Vilar's typology.

Considering the error types, the one-to-one equivalent error types Accuracy> Mistranslation in DQF-MQM and Wrong lexical Choice in Vilar's, the annotations for both typologies follow the same trend as most of the annotations were annotated on the environmental statistical translations. Concerning Accuracy> Typography and Punctuation error types, the maximum of errors appears on the environmental neural translations for both typologies.

However, when one DQF-MQM error type corresponds to two Vilar's ones, the results are not fully equivalent.
For the Fluency> Word Order and Word Order> local range and long range error types, it is the same trend for the patent neural translation as no error was annotated for those types. But on the patent and environmental statistical translations, while the same number of errors were annotated under DQF-MQM, it is not the case for Vilar's.
We obtain the same figures for Accuracy> Omission and Missing Words> Filler Word and Content words, as these error types were used only on the environmental translations, but more of them were annotated on Robert than on Jude Law under DQF-MQM, while their amount is equivalent for Madagascar and New Zealand under Vilar's.

Similarly, when two DQF-MQM error types correspond to one single error type in Vilar's, the results are slightly different. Looking at Fluency> Word Form> POS and Agreement and Incorrect Word> Incorrect form, more errors were annotated on statistical translations on both domains under both typologies, but most of the errors were annotated on George under DQF-MQM while most of them were annotated on New Zealand under Vilar's.

The only contradiction occurs with Accuracy> Addition and Extra Words error types, as no error of this type was annotated for the patent translations under DQF-MQM while it was under Vilar's. The same annotations happen on the environmental translations under both typologies.
We then assume that having two different typologies implies annotating in a different way the same errors.

3.2 Qualitative results

To explain those figures, we studied how the data were annotated using the two different typologies.
We have found too main discrepancies. A recurring discrepancy is the selection issue: difference in the number of annotations depending on the typology used. The second recurring issue is the annotation issue: for the same error, the annotator annotated it under one of the typologies and not the other one.

3.2.1 Selection issues

In the corpus, we found out eight occurrences of selection issues. In this section, we explicit five examples.
In the first example below, taken from the environmental statistical translations, "this" translated by "la présente", was annotated as Accuracy>Mistranslation (DQF-MQM), counting for one single error, while under Vilar's, the annotator divided it into two errors: "la" as an Incorrect words>Sense>Wrong lexical choice error, and "présente" as an Incorrect words> Extra Words error. Consequently, we have got one more error with Vilar's typology than with the DQF-MQM. The annotator did not cut the translated phrase in the same manner, given way to one more annotated error under Vilar's.

Example 1:
GB: Our sustainable prosperity will depend on this.
FR: *Notre prospérité durable dépendra de la présente.

In our second example from the same translation corpus, "action" was translated into "la lutte contre le changement". The annotator considered "action" as not translated into French, assigning the Accurracy> Omission error type (DQF-MQM,) while "la lute contre le changement" was annotated under Accuracy> Addition. Contrastingly, the annotator annotated "la lutte contre le changement" under Incorrect Word> Sense> Wrong Lexical Choice type (Vilar's). Two errors were annotated under DQF-MQM versus only one for Vilar's. The types are not equivalent under both typologies. Also, it is the only time that an error labeled Accuracy> Addition under DQF-MQM) is not annotated into its equivalent, i.e. Incorrect Words> Extra Words under Vilar's.
Again, we do not have the same cut and no equivalence between the error types.

Example 2:
GB: […] lessons learned in climate action with officials from other countries […]
FR: *[…] les enseignements tirés dans la lutte contre le changement climatique avec des représentants d'autres pays […]

In another example from the environmental neural translation, "acting on" was not translated. Under DQF-MQM, the annotator used the two error types on omission, i.e. Accuracy> Omission and Fluency> Grammar> Function Words – Missing, to annotate respectively "acting" as one error and "on" as another one, resulting in two errors. On the contrary, under Vilar's, the annotator kept "acting on" as a single error and annotated it under Missing Words> Filler Words. We thus wonder if this discrepancy is due to a lack of consistency from the annotator, or a bias due to the error type labelling in the two typologies.

Example 3:
GB: […] protecting the environment and acting on climate change must go together.
FR: *[…] la protection de l'environnement et le changement climatique doivent aller de pair.

In our next example from the patent neural translations, "at" was translated into "au niveau d'" and annotated as Accuracy> Over-translation. Under Vilar's, "niveau d'" was considered as Incorrect> Extra Words. We consequently, have only one error but we do not have the same cut nor an equivalent error type for both typologies.

Example 4:
GB: […] a longitudinally extending seat extension attached at a post end […]
FR: *[…] une extension de siège s'étendant longitudinalement fixée au niveau d'une extrémité de montant […]

In our last example, "linkages" was considered as an Accuracy> Mistranslation (DQF-MQM) while "of" was considered as a Missing Words> Content Words (Vilar's). We can wonder if the complexity of an error entails a bias on the annotation whatever the typology, thus, asking to ourselves if it exists a threshold of the number of errors above which an annotator cannot annotate properly.

Example 5 :
GB: […] including carbon pricing, linkages of carbon market policies […]
FR: *[…] y compris la tarification du carbone, des liens politiques du marché du carbone […]

3.2.2 Annotation issues

Concerning the errors that were annotated under one typology and not the other one, we found out 18 occurrences. 94% of those errors were annotated under Vilar's Typology and not DQF-MQM opposed to 6% annotated under DQF-MQM typology and not Vilar's one. We have selected 4 examples to illustrate this issue.

Considering example 1 below, taken from the environmental statistical translations, "ambitious" has been translated into "ambitieux" (masc.) instead of "ambitieuse" (fem.). The annotator has not annotated it as an error under Vilar's but has under DQF-MQM assigning Incorrect Words> Incorrect Form. So again, we can wonder how many mistakes an annotator can handle at once while annotating.

Example 1:
GB: […] and engage with sub-national governments and non-state actors to develop ambitious climate solutions.
FR: *[…] et de dialoguer avec les gouvernements infranationaux et les acteurs non étatiques à mettre au point des solutions climatiques ambitieux.

In our next example taken from the patent statistical translations, "method" was translated under "un procédé" and the annotator has considered it as an Incorrect Words> Style error (Vilar's) while she has not annotated it under DQF-MQM.

Example 2:
GB: method enables website providers to prohibit non-human users […]
FR: *Un procédé permet aux fournisseurs de site web pour interdire des utilisateurs non humains […]

In the following example, taken again from the patent statistical translations, "complete" was translated into "terminer" and was annotated under Vilar's as Incorrect Words> Style error while not annotated under DQF-MQM.
As only Vilar's typology offers the error type Incorrect Words> Style, our annotator has no equivalent in DQF-MQM, implying a bias in the annotations.

Example 3:
GB: […] an instruction provided to the user on how to complete the authentication challenge.
FR: *[…] une instruction fournie à l'utilisateur sur la manière de terminer le défi d'authentification.

In our last example, from the environmental statistical translations, "recognised" was translated as "admis" and annotated as Incorrect Words> Sense> Wrong Lexical Choice under Vilar's and not under DQF-MQM, while it exists an equivalent error type is this typology.

Example 4:

GB: Canada has long recognized the principle that free and open trade […]
FR: *Le Canada a admis depuis longtemps le principe selon lequel des échanges commerciaux libres et ouverts[…]
Recognized admis Incorrect Words> Sense> Wrong Lexical Choice

4. Discussion and further work

We have seen that the typology choice impacts on how the error occurrences are selected as well as on how the types are selected.
The typology choice also impacts on the number of errors, thus on metrics relying on them. It also opens a way to further studies on the impact of evaluation criteria on quality assessment. We would like to investigate further in that way by analysing the annotations from more annotators, and not only at the sentence level but at the document level. And we would also work on the threshold on the number of errors above which an annotator cannot cognitively annotate properly a translation.

References

Bregeon, Dany, and Jean-Yves Antoine, Jeanne Villaneau, Anaïs Lefeuvre-Halftermeyer. Redonner du sens à l'accord interannotateurs : vers une interprétation des mesures d'accord en termes de reproductibilité de l'annotation. *Traitement Automatique des Langues*, ATALA, 2019, 60 (2), pp.23. ⟨hal-02375240⟩

Eckart, Thomas and Uwe Quasthoff, Dirk, Goldhahn. (2012). The Influence of Corpus Quality on Statistical Measurements on Language Resources. May 2012. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), 2012

Esperança-Rodier, Emmanuelle and Francis, Brunet-Manquat, and Sophia Eady. 2019. ACCOLÉ: A Collaborative Platform of Error Annotation for Aligned Corpora. In Translating and the computer 41.

Klubicka Filip, and Antonio, Toral, Victor M. Sánchez-Cartagena. 2017. Fine-grained human evaluation of neural versus phrase-based machine translation. *Prague Bull Math Linguist* 108(1):121–132

Koehn, Philippan and Rebecca, Knowles. 2017. Six Challenges for Neural Machine. On *Translation Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, Canada, August 4, 2017. c 2017 Association for Computational Linguistics

Linzen, Tal. ACL 2020. How Can We Accelerate Progress Towards Human-like Linguistic Generalization? arXiv pre-print:2005.00955

Lommel, Arle, and Alan, K. Melby. 2018. Tutorial: MQM-DQF: A Good Marriage (Translation Quality for the 21st Century). In *Proceedings of 13th Conference of the Association for Machine Translation in the Americas (Volume 2: User Papers). Vol. 2.*

Mathet, Yann, and Antoine Widlöcher, Karen Fort, Claire François, Olivier Galibert, et al.. Manual Corpus Annotation: Giving Meaning to the Evaluation Metrics. International

Conference on Computational Linguistics, Dec 2012, Mumbaï, India. pp.809–818. ffhal-00769639f

Popović, Maja. Error Classification and Analysis for Machine Translation Quality Assessment Published in Moorkens, Joss, Sheila Castilho, Federico Gaspari, and Stephen Doherty (Eds). *Translation Quality Assessment Machine Translation: Technologies and Applications 1, 2018 https://doi.org/10.1007/978-3-319-91241-7_7*

Vilar, David and Jia Xu, Luis Fernando D'Haro and al. 2006. Error analysis of statistical machine translation output. In the *Proceedings of 5th International Conference on Language Resources and Evaluation.* Pages 97-702.

ANNEXES

ANNEX 1 – Detailed Corpora

Document Type	**Segments**	**Source words**	**MT**	**Target words**	**DQF-MQM**		**Vilar**	
Patent	9	272	WIPO Translate N	305	Brad Pitt		Hawaii	
					11	annotations	13	annotations
			WIPO Translate S	302	George Clooney		Madagascar	
					25	annotations	29	annotations
Sub-Total	9	272	-	607	36	annotations	42	annotations
European Community Document on Climate Change	28	519	eTranslation	768	Jude Law		Nouvelle Calédonie	
					35	annotations	36	annotations
			legacy MT@EC	776	Robert		New Zealand	
					51	annotations	59	annotations
Sub-Total	28	519	-	1544	86	annotations	95	annotations
TOTAL	**37**	**791**	**-**	**2151**	**122**	**annotations**	**137**	**annotations**

ANNEX 2 - Ranking of error types for both typologies according to the percentage of errors annotated on the 4 documents

Rank	DQF-MQM		VILAR	
1	Accuracy> Mistranslation	38%	Incorrect Words> Sense> Wrong Lexical Choice	40%
2	Accuracy> Addition	16%	Incorrect Words> Extra Words	17%
3	Fluency> Grammar> Word order	9%	Incorrect Words> Incorrect Form	15%
4	Accuracy > Omission	7%	Word Order> Word Level> Local Range	7%
5	Fluency> Grammar> Word form - Part of Speech	7%	Punctuation	6%
			Incorrect Words> Style	6%
6	Fluency > Typography	6%	Missing Words> Filler Words	4%
	Fluency> Grammar> Word form – Agreement	6%		
7	Accuracy> Over-translation	5%	Incorrect Words> Sense> Incorrect Desambiguation	2%
			Word Order> Phrase Level> Local Range	2%
8	Fluency> Grammar > Function Words – Incorrect	3%	Missing Words> Content Words	1%
9	Accuracy> Mistranslation> Overly literal	1%		
	Fluency> Grammar> Function words – Extraneous	1%		
	Fluency> Grammar > Function Words – Missing	1%		
	Fluency> Grammar > Word form – Tense	1%		

ANNEX 3 – Number and percentage of annotations per document using DQF-MQM annotation typology

DQF-MQM	Brad Pitt		George Clooney		Jude Law		Robert		Total		Ranking
Accuracy> Mistranslation	5	10,87%	11	23,91%	12	26,09%	18	39,13%	46	38%	1
Accuracy> Addition	0	0,00%	0	0,00%	8	42,11%	11	57,89%	19	16%	2
Fluency> Grammar> Word order	0	0,00%	5	45,45%	1	9,09%	5	45,45%	11	9%	3
Accuracy> Omission	0	0,00%	0	0,00%	4	44,44%	5	55,56%	9	7%	4
Fl> Grammar> Word form- POS	2	25,00%	3	37,50%	1	12,50%	2	25,00%	8	7%	5
Fluency> Typography	1	14,29%	0	0,00%	4	57,14%	2	28,57%	7	6%	6
Fl> Grammar> Word form-Agree	1	14,29%	4	57,14%	0	0,00%	2	28,57%	7	6%	6
Accuracy> Over-translation	2	33,33%	0	0,00%	3	50,00%	1	16,67%	6	5%	7
Fl> Grammar> Function Words- Inc	0	0,00%	1	25,00%	0	0,00%	3	75,00%	4	3%	8
Ac> Mistranslation> Overly literal	0	0,00%	0	0,00%	0	0,00%	1	100,00%	1	1%	9
Fl> Grammar> Function words- Extr	0	0,00%	1	100,00%	0	0,00%	0	0,00%	1	1%	9
Fl> Grammar> Function Words- Miss	0	0,00%	0	0,00%	1	100,00%	0	0,00%	1	1%	9
Fl> Grammar> Word form - Tense	0	0,00%	0	0,00%	0	0,00%	1	100,00%	1	1%	9
Total	**11**	**9,09%**	**25**	**20,66%**	**34**	**28,10%**	**51**	**42,15%**	**121**	**100%**	

ANNEX 4 – Number of annotations and percentage per document using Vilar's annotation typology

VILAR	Hawaii		Madagascar		Nouvelle Calédonie		New Zealand		Total		Ranking
Incorrect Words > Sense > Wrong Lexical Choice	5	9,09%	13	23,64%	13	23,64%	24	43,64%	55	40%	1
Incorrect Words > Extra Words	2	8,70%	1	4,35%	9	39,13%	11	47,83%	23	17%	2
Incorrect Words > Incorrect Form	3	15,00%	7	35,00%	1	5,00%	9	45,00%	20	15%	3
Word Order > Word Level > Local Range	0	0,00%	5	55,56%	0	0,00%	4	44,44%	9	7%	4
Punctuation	1	12,50%	1	12,50%	4	50,00%	2	25,00%	8	6%	5
Incorrect Words > Style	2	25,00%	2	25,00%	3	37,50%	1	12,50%	8	6%	5
Missing Words > Filler Words	0	0,00%	0	0,00%	4	66,67%	2	33,33%	6	4%	6
Incorrect Words > Sense > Incorrect Desambiguation	0	0,00%	0	0,00%	1	33,33%	2	66,67%	3	2%	7
Word Order > Phrase Level > Local Range	0	0,00%	0	0,00%	1	33,33%	2	66,67%	3	2%	7
Missing Words > Content Words	0	0,00%	0	0,00%	0	0,00%	2	100,00%	2	1%	8
Total	**13**	**9,49%**	**29**	**21,17%**	**36**	**26,28%**	**59**	**43,07%**	**137**	**100%**	