Document-Level Machine Translation Evaluation Project: Methodology, Effort and Inter-Annotator Agreement

Final Report
28/01/2021

Sheila Castilho
ADAPT Centre, School of Computing, Dublin City University
sheila.castilho@adaptcentre.ie

1 Introduction

This project is funded by the European Association for Machine Translation through its 2020 sponsorship of activities programme. This document contains the final report of the project, which aims at identifying the complexity when assigning a single score to full texts and investigate the difference in IAA between sentence and document level evaluation set-ups.

2 Progress and Milestones Achieved in the Project

2.1 Phase I:

The first phase of the project conducted an evaluation study with four professional translators and added another one at later stage. The translators evaluated (1) fluency/adequacy, and error mark-up in the PET tool (Aziz et al., 2012); and (2) pairwise ranking in Google spreadsheet. The evaluation was carried in two scenarios: (A) evaluation at the sentence level, showing randomised sentences, one at a time, and (B) evaluation at a document level. After each scenario was complete, translators answered a post-task questionnaire about the evaluation. In total, 1000 sentences were evaluated by each translator, 500 in scenario A (test set 1) and 500 in scenario B (test set 2). We observed that one of the translators (T1) reported to have a difficult time to hand in the evaluation (due to COVID-19) and self-reported he was distracted while doing it. Therefore, we decided to add a 5th translator (T5) to be compared with T1&T2 in order to get a further understanding of issues observed. For results with T1&T2&T5 we report Fleiss Kappa.

2.1.1 Results

Agreement – inter-annotator agreement IAA for all the assessments have been calculated with Cohen's Kappa (weighted and non-weighted) and Fleiss Kappa, as well as inter-rater reliability (IRR). Results are showed in Table 1. For TEST SET 1, higher IAA is seen for adequacy in sentence-level set up, but higher K for document-level for the weighted variation. Higher IAA is seen for fluency in document-level set up. Ranking shows higher IAA for sentence-level set up. Error annotation

			Adequacy	Fluency	Ranking	Error	
						binary	type
TEST SET 1 sentence	Fleiss Kappa		0.05	0.88	0.26	0.168	0.02
	IRR		67%	63%	59%	60%	56%
TEST SET 1 document	Kappa	NW	0.01	0.41	0.22	n/a	0.31
		W	0.23	0.25			
	Pearson		*0.64*	*0.73*	*0.36*	*0.08*	
	p-value		*0*	*0*	*0.04*	*0.49*	
	IRR		44%	56%	56%	100%	53%
			Adequacy	Fluency	Ranking	Error	
						binary	type
TEST SET 2 sentence	Kappa	NW	0.34	0.27	0.29	0.49	0.38
		W	0.41	0.34			
	Pearson		*0.53*	*0.42*	0.41	0.7	
	p-value		*0*	*0*	0	0	
	IRR		63%	57%	53%	76%	56%
TEST SET 2 document	Fleiss Kappa		-0.12	-0.12	0.145	-0.079	-0.018
	IRR		42%	50%	47%	88%	50%

Table 1- Inter-annotator agreement for all assessments in phase I

shows higher K for the document-level set up but higher IRR for sentence-level.[1][2] Results for TEST SET 2 were convoluted before adding T1&T2, and after adding T5 the differences between the sentence-level and the document-level set ups are still larger, with K reaching negative scores.

Perceived effort (Post-task Questionnaire) – Results from the post-task questionnaire (Table 2) with all 5 translators' assessments suggest that while translators prefer to see full texts than single sentences, they would rather see sentence pairs and paragraphs than having to assess full documents. As well, they find assessing a full document more tiring than the alternative.

Questions	scale	single sentence	full texts
I was *always* able to understand the meaning of the source [sentence/texts]	1 disagree- 6 agree	5	5.4
I was *always* able to understand the meaning of the translated [single sentence/full texts]	1 disagree- 6 agree	4.2	3.8
I was *always* able to recognise all the problems with the translation of [single sentence/full texts]	1 disagree- 6 agree	5.2	4.8
I would have preferred to evaluate [full texts/single sentences] than [single sentence/full texts]	1 disagree- 6 agree	4	4.6
I would have preferred to evaluate pair of sentences than [single sentence/full texts]	1 disagree- 6 agree	3.8	5
I would have preferred to evaluate full paragraphs than [single sentence/full texts]	1 disagree- 6 agree	3.6	4.2
I was satisfied with the evaluation I provided for the [single sentence/full texts] job	1 disagree- 6 agree	4.8	5
Spotting errors in the each translated [single sentence/full texts] was	1 very difficult - 6 very easy	5.2	4.4
Assessing the translation quality on a [single sentence/full texts] level was:	1 very difficult - 6 very easy	4.6	4.2
Assessing the translation quality on a [single sentence/full level] was:	1 very tiring - 6 not at all	3.2	1.8

Table 2- Post-task questionnaire with judgements for perceived effort in phase I

2.2 Phase II:

In phase II, a document-level evaluation was performed in two scenarios:
A) sentence-level: with translators, giving one score per random sentence.
B) document-level: with translators giving one score per sentence while having access to the full text. Translators also gave a general score for the full text.[3]

This is consistent with the responses to the post-task questionnaire in Phase I showed in Table II and with the context-span necessary for translation as seen in Castilho et al. (2020). For this phase, the number of sentences was reduced and the number of translators increased. In total, 14 short documents (513 sentences) from various sources were used: WMT newstest 2019, OPUS Corpus Ted Talk, excerpts from books,[4] and product reviews.[5] Eight translators participated in Phase II, and similarly to Phase I, they were grouped in such a way that the same translator would not assess the same sentence/document twice. After assessing the sentences/documents, translators were asked to fill in a post-task questionnaire.

2.2 Results:

We calculated Fleiss Kappa, IRR and Krippendorff's Alpha Reliability Estimate to compare the IAA between random sentence-level and document-level scenarios. Results in Table 3 show that for test set 1,

[1] Note that Kappa for binary error in the document-level set up is 1 (100% agreement per IRR) as translators agreed that all documents contained at least one error. However, Kappa penalises it as all the ratings fall into a single category.
[2] Error mark-up was divided into *binary* (when raters agree whether there was an error (any type) or no errors in the sentence/document) and *type* (when raters agree on the exact error type found in the sentence/document).
[3] The results of that evaluation is still being processed. We intend to submit a journal article with all the results soon (see section 5)
[4] The excerpts from *The Girl on the Train* and *The Fault in Our Stars* were found freely available online.
[5] Product reviews were collect on the amazon.com website.

sentence-level scenario shows a higher IAA for all metrics compared to the document-level one. However, it is interesting to note that the difference is not as distinct as it was in phase I. Moreover, results of test set 2 shows higher IAA for fluency and ranking for the document-level scenario. Adequacy and error mark-up are slightly higher for the sentence-level scenario but the difference is quite small. These results confirm that document-level evaluation that yields better annotator agreement is the one translators give a score per sentence and have access to the full text.

		Adequacy	Fluency	Ranking	Error	
					binary	type
TEST SET 1 sentence	Fleiss Kappa	0.320	0.288	0.401	0.288	0.279
	p.value	0.00	0.00	0.00	0.00	0.00
	Krippendorff's Alpha	0.321	0.289	0.402	0.288	0.279
	IRR	70%	69%	61%	68%	65%
TEST SET 1 document	Fleiss Kappa	0.294	0.162	0.363	0.223	0.247
	p.value	0.00	0.00	0.00	0.00	0.00
	Krippendorff's Alpha	0.295	0.162	0.363	0.224	0.247
	IRR	55%	49%	58%	63%	55%

		Adequacy	Fluency	Ranking	Error	
					binary	type
TEST SET 2 sentence	Fleiss Kappa	0.236	0.164	0.384	0.273	0.256
	p.value	0.00	0.00	0.00	0.00	0.00
	Krippendorff's Alpha	0.237	0.165	0.385	0.269	0.249
	IRR	59%	56%	60%	62%	58%
TEST SET 2 document	Fleiss Kappa	0.223	0.198	0.428	0.150	0.163
	p.value	0.00	0.00	0.00	0.00	0.00
	Krippendorff's Alpha	0.224	0.199	0.432	0.1510	0.166
	IRR	59%	61%	62%	60%	55%

Table 3 - Inter-annotator agreement for all assessments in phase II.

Perceived effort (Post-task Questionnaire) – Results from the post-task questionnaire in Table 4 show that translators preferred the document-level scenario for evaluating the quality of the machine translation system. They found this methodology easier to understand both source and target, recognise errors and choosing the better of two translations. Moreover, translator were more confident in their assessment when judging the translation with the document-level methodology.

3 Budget

For phase II, the budget from EAMT was 3000 euros (second instalment). Each translator was paid 350 euros for the task, totalling 2.800 euros (350x8). The reminder 200 euros was used to pay T5 to re-assess the same test set as T1 in phase I.[6]

[6] T5 is a close colleague of mine and accepted the low rate payment to perform the evaluation as a personal favour.

Questions	scale	random single sentences	full texts
Understanding the meaning of the random SOURCE [the random sentences/in each sentence, with access to the full document] in general was	1 very difficult - 6 very easy	4.37	5.75
Understanding the meaning of the random TRANSLATED [the random sentences/in each sentence, with access to the full document] in general was	1 very difficult - 6 very easy	3.87	5.12
Recognising the ADEQUACY problems in [the random sentences/in each sentence, with access to the full document] in general was	1 very difficult - 6 very easy	4.12	5.25
Recognising FLUENCY problems in [the random sentences/in each sentence, with access to the full document] in general was	1 very difficult - 6 very easy	4.62	4.87
Spotting ERRORs in each of [the random sentences/in each sentence, with access to the full document] in general was	1 very difficult - 6 very easy	4.5	5.12
Chosing the best of two translations in a random sentence evaluation was	1 very difficult - 6 very easy	4.12	4.87
In general, assessing the translation quality on a [sentence/document] level was:	1 very difficult - 6 very easy	4	5
For me, assessing the translation quality on a [sentence/document] level was:	1 very tiring - 6 not at all	3.75	4.62
I was confident with every assessment I provided for the [sentence/document] level evaluation tasks	1 strongly agree - 6 strongly disagree	4.12	4.62
I could have done a more accurate assessment if [I had had access to the full text/was assessing random sentences]	1 strongly agree - 6 strongly disagree	5.12	1.37

Table 4 - Post-task questionnaire with judgements for perceived effort in phase II

4 Next Steps

We intend to analyse the results of this experiment from all points of view in order to get a better understanding of the methodology. Moreover, these experiments performed here will serve as a pilot for a bigger project - named DELA Project[7] - which I will be conducting in the next 2 years. This project aims at analysing new methodologies for document-level machine translation evaluation with both human and automatic metrics.

5 Publications

Sheila Castilho. On the same page? Comparing inter-annotator agreement in sentence and document level human machine translation evaluation. In the Fifth Conference on Machine Translation, WMT'20, November 2020.

Sheila Castilho. Document-Level Machine Translation Evaluation Project: Methodology, Effort and Inter-Annotator Agreement. In the 22nd Annual Conference of the European Association for Machine Translation, EAMT'20, November 2020

I intend to submit a paper with the results of Phase II to the Transactions of the Association for Computational Linguistics (TACL) journal.

[7] https://adaptcentre.ie/projects/dela-project/