Engaging Content
Engaging People

A World Leading SFI Research Centre

On the Same Page? Comparing IAA in Sentence and Document Level Human MT Evaluation

Sheila Castilho

sheila.castilho@adaptcentre.ie

@_SheilaCastilho

Why document-level MTEval?

- Potential problems with random single-sentence evaluation

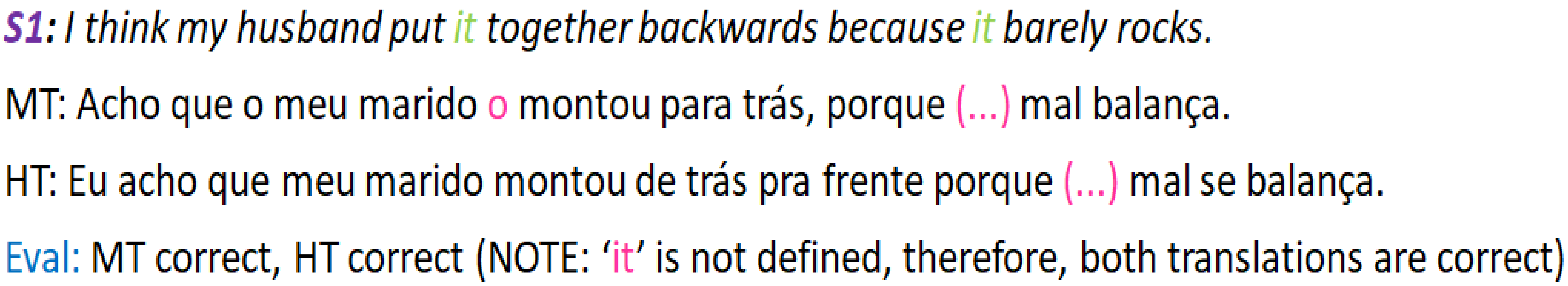
Sent 1	I put it in my car	What is “it”?
Sent 2	“Speak English!”- said the Eaglet	Who that order was addressed to?
Sent 3	Yes, she did.	What did she do?
Sent 4	I am lonely.	Is “I” feminine or masculine – need to know to translate “lonely”. Is “am” a transitory or an inherent characteristic of “I”? Ser/estar

Why document-level MTEval?

Castilho, Popovic and Way (2020)

- context span for reviews, subtitles, and literature domains
- 33% require more context than the sentence itself to be translated
  - 23% required more than two previous sentences to be properly translated.
- Common issues: ambiguity, terminology, and gender agreement were the most found to hinder translation.

Potential Problems with random sentence evaluation



***S1**: I think my husband put it together backwards because it barely rocks.*

MT: Acho que o meu marido o montou para trás, porque (...) mal balança.

HT: Eu acho que meu marido montou de trás pra frente porque (...) mal se balança.

Eval: MT correct, HT correct (NOTE: ‘it’ is not defined, therefore, both translations are correct)

- Both the MT and human translations (HT) are correct according to the evaluation (Eval), as it is impossible to determine the gender of the pronoun 'it'. While MT chooses to translate “it” into masculine (o), HT did not add a pronoun

A

S1: I think my husband put it together backwards because it barely rocks.

MT: Acho que o meu marido o pôs para trás, porque (...) mal balança.

HT: Eu acho que meu marido montou de trás pra frente porque (...) mal se balança.

Eval: MT correct. HT correct (NOTE: ‘it’ is not defined, therefore, both translations are correct)

B

1+S1: It is pretty. I think my husband put it together backwards because it barely rocks.

MT: (...) É bonito. Acho que o meu marido o pôs para trás, porque (...) mal balança.

HT: (...) É/está bonito/a. Eu acho que meu marido montou de trás pra frente porque (...) mal se balança. (NOTE: The gender for “it” is necessary to choose between ele/ela; and more contexto needed to know which verb to use for “is”)

Eval: MT correct. HT correct. (NOTE: ‘it’ is not defined, therefore, both translations are correct)

C

2+S1: It is more of an olive green than a mint green. It is pretty. I think my husband put it together backwards because it barely rocks.

MT: (...) É mais verde-oliva do que verde-menta. (...) É bonito. Acho que o meu marido o pôs para trás, porque (...) mal balança.

HT: (...) É/está mais pra um verde-oliva do que um verde-menta. (...) É/está bonito/a. Eu acho que meu marido montou de trás pra frente porque (...) mal se balança. (note: I need to know the gender for “it”, to choose between ele/ela; and need to know which verb to use for “is”)

Eval: MT correct. HT correct. (NOTE: ‘it’ is not defined, therefore, both translations are fluent)

D

3+S1: This chair is way darker than it is in the picture. It is more of an olive green than a mint green. It is pretty. I think my husband put it together backwards because it barely rocks.

MT: Esta **cadeira** é muito xxx escura do que está na imagem. (...) É mais verde-oliva do que verde-menta. (...) É bonito. Acho que o meu marido o pôs para trás, porque (...) mal balança.

HT: Esta **cadeira** é muito mais escura do que na imagem. Ela está mais pra um verde-oliva do que um verde-menta. É bonita. Eu acho que meu marido montou de trás pra frente porque ela mal se balança.

Eval: MT incorrect – no agreement with gender, missing word, mistranslation, and low fluency. HT correct.

Neither A, B, and C solves the problem, since “it” is still not defined.

Only in D (3+S1) is that we can identify what “it” is (chair=female), and therefore evaluate the sentence properly.

Document-level Eval methodology

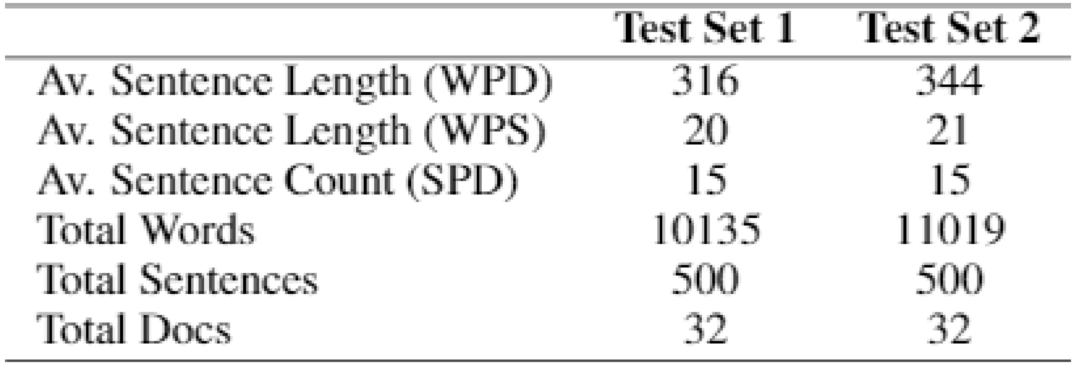
- **Läubli, Sennrich and Volk (2018):**
  - Professional translators
  - Fluency and Adequacy ranking
  - One score per text
- **Toral, Castilho, Hu and Way (2018):**
  - Professional translators and bilinguals
  - consecutive single sentences
  - ranking
- **WMT (2019):**
  - Direct Assessments (accuracy) with bilinguals
    a) sentence-score+document level evaluation
    b) document-score+document-level evaluation

In this paper

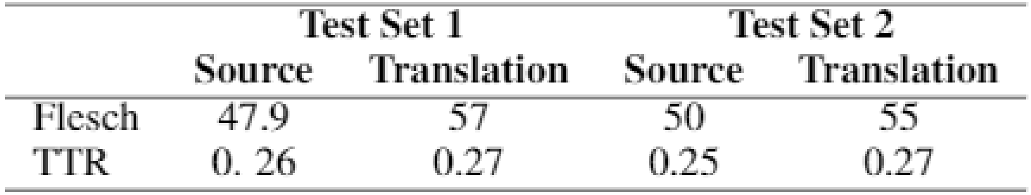
- Comparison of IAA between (A) sentence-level and (B) document-level set-ups
  - Kappa (W, NW, F, 0-1)and IRR(%)
    - Fluency, adequacy
    - Error mark-up
    - Pairwise ranking
- Effort to perform the task
  - Self-assessment
    - Post-task Questionnaire
  - Time
    - Time logged in the PET tool

Methodology

- Corpus
  - WMT *newstest2019*
  - 64 full texts (1K sentences) divided into 2 test sets:
    - 32 full texts (500 sentences) per scenario
  - Both test sets and translations are comparable
  - Translated from EN into PT-BR with Google Translate (for adequacy, fluency, and error mark-up) and also with DeepL for the ranking pairwise comparison



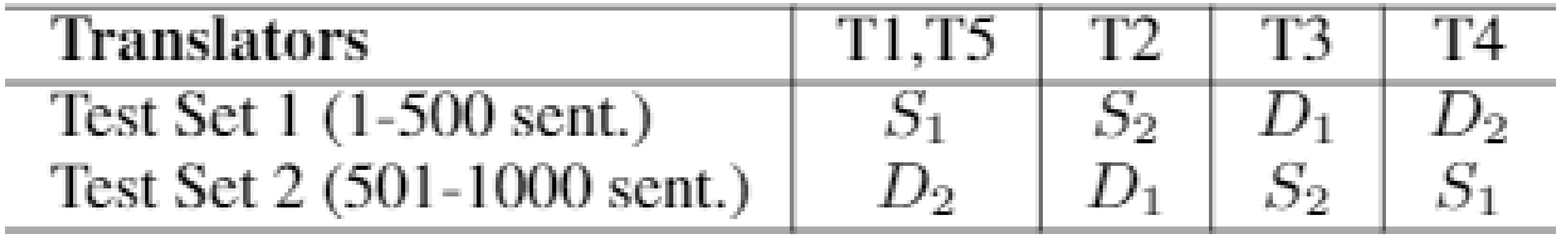
	Test Set 1	Test Set 2
Av. Sentence Length (WPD)	316	344
Av. Sentence Length (WPS)	20	21
Av. Sentence Count (SPD)	15	15
Total Words	10135	11019
Total Sentences	500	500
Total Docs	32	32



	Test Set 1		Test Set 2	
	Source	Translation	Source	Translation
Flesch	47.9	57	50	55
TTR	0. 26	0.27	0.25	0.27

Methodology

- Four professional translators (EN → PT-br) (1 added at a later stage)



Translators	T1,T5	T2	T3	T4
Test Set 1 (1-500 sent.)	S_1	S_2	D_1	D_2
Test Set 2 (501-1000 sent.)	D_2	D_1	S_2	S_1

Methodology

- PET Tool (Aziz et al 2012)
  - **Adequacy**
    “How much of the meaning expressed in the source appears in the translation?”
    1. None of it, 2. Little of it, 3. Most of it, 4. All of it
  - **Fluency**
    “How fluent was the translation?”
    1. No fluency, 2. Little fluency, 3. Near native, 4. Native

1 score per sentence *VS* 1 score per document

  - **Error Markup**

    No issues, Mistranslation, Untranslated, Word Form, and Word order.

sentence tagging *VS* document tagging

Methodology

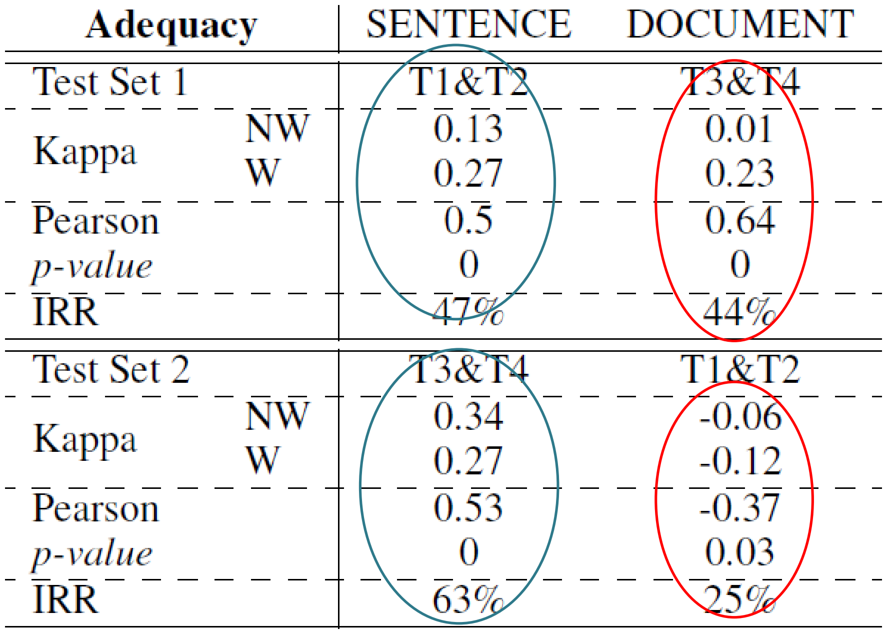
- Spreadsheet
  - Pairwise ranking
    - Google vs DeepL

Best of two translated sentences *VS* Best of two translated documents

- Post-task Questionnaire
  - Preferences
  - Satisfaction
  - Level of difficulty
  - Level of effort

Results - IAA

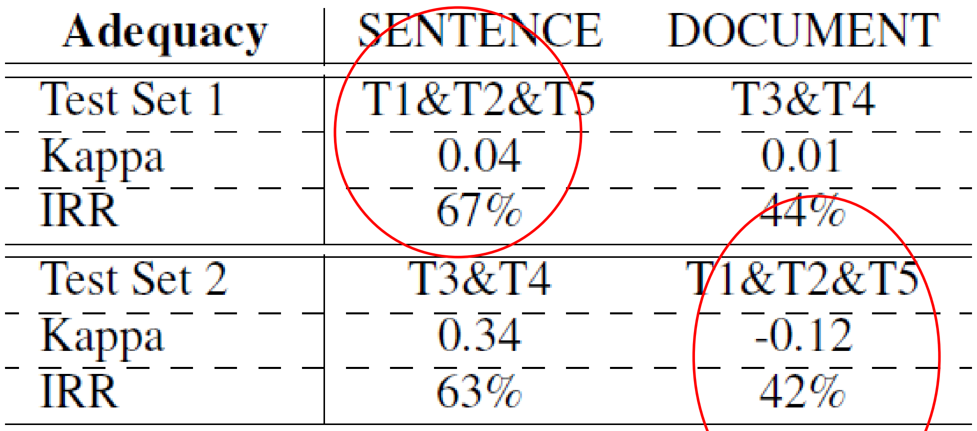
- Adequacy



Adequacy		SENTENCE	DOCUMENT
Test Set 1		T1&T2	T3&T4
Kappa	NW	0.13	0.01
	W	0.27	0.23
Pearson		0.5	0.64
p-value		0	0
IRR		47%	44%
Test Set 2		T3&T4	T1&T2
Kappa	NW	0.34	-0.06
	W	0.27	-0.12
Pearson		0.53	-0.37
p-value		0	0.03
IRR		63%	25%

- Test Set 1: IAA higher score for single-sentence scenario.
- Test Set 2: very discrepant scores – negative *K* scores for document level

- T5: decline in *K* for both scenarios and increase in IRR

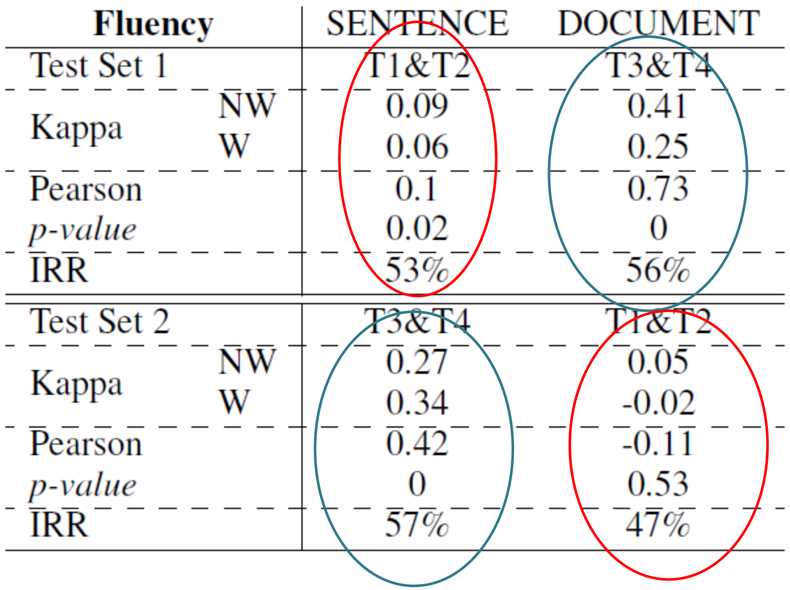


Adequacy	SENTENCE	DOCUMENT
Test Set 1	T1&T2&T5	T3&T4
Kappa	0.04	0.01
IRR	67%	44%
Test Set 2	T3&T4	T1&T2&T5
Kappa	0.34	-0.12
IRR	63%	42%

IAA is higher in the sentence-level scenario

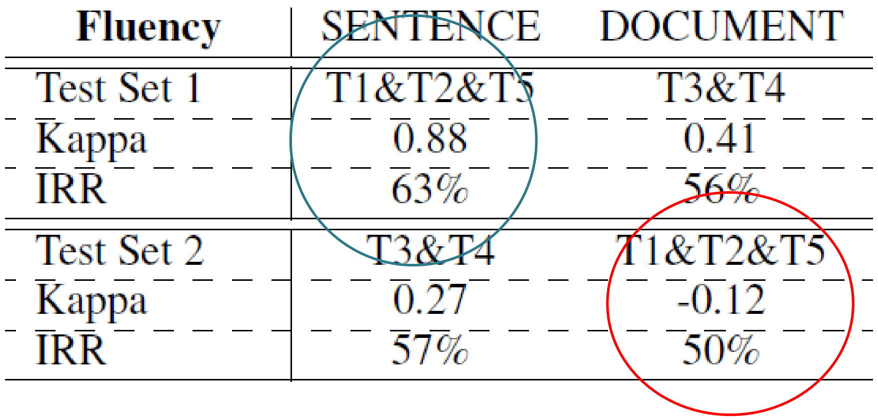
Results - IAA

- Fluency



Fluency		SENTENCE	DOCUMENT
Test Set 1		T1&T2	T3&T4
Kappa	NW	0.09	0.41
	W	0.06	0.25
Pearson		0.1	0.73
p-value		0.02	0
IRR		53%	56%
Test Set 2		T3&T4	T1&T2
Kappa	NW	0.27	0.05
	W	0.34	-0.02
Pearson		0.42	-0.11
p-value		0	0.53
IRR		57%	47%

- Test Set 1: higher IAA scores for document-level scenario
- Test Set 2: Low IAA scores (negative *K*) for document level



Fluency	SENTENCE	DOCUMENT
Test Set 1	T1&T2&T5	T3&T4
Kappa	0.88	0.41
IRR	63%	56%
Test Set 2	T3&T4	T1&T2&T5
Kappa	0.27	-0.12
IRR	57%	50%

- T5: increase in IAA for sentence-level scenario
- Decrease in *K* for document-level

IAA is higher in the sentence-level scenario, but fluency assessment benefit from full documents

Results - IAA

Error

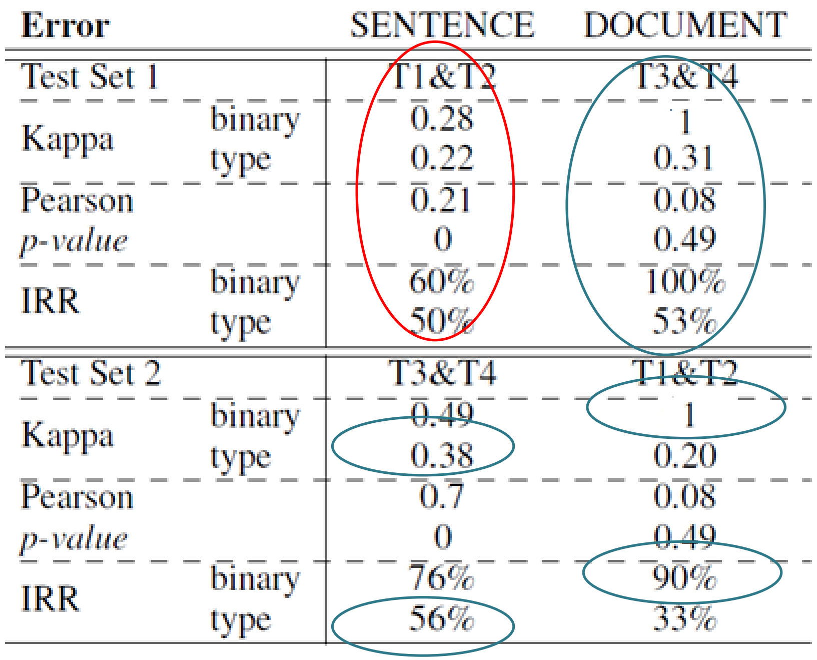
- Simple taxonomy – (agreement between translators)

- Each sentence or document could be annotated with more than one error category, but each error category could be assigned only once

- results were divided into:
  - Binary - when raters agree whether there was an error (any type) or no errors in the sentence/document,
  - Type - when raters agree on the exact error type found in the sentence/document.

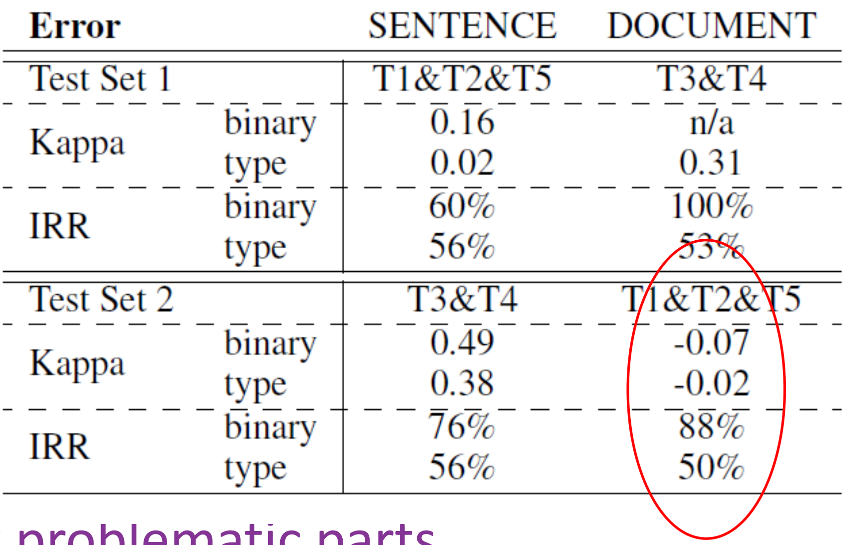
Results - IAA

- Error



Error		SENTENCE	DOCUMENT
Test Set 1		T1&T2	T3&T4
Kappa	binary	0.28	1
	type	0.22	0.31
Pearson		0.21	0.08
p-value		0	0.49
IRR	binary	60%	100%
	type	50%	53%
Test Set 2		T3&T4	T1&T2
Kappa	binary	0.49	1
	type	0.38	0.20
Pearson		0.7	0.08
p-value		0	0.49
IRR	binary	76%	90%
	type	56%	33%

- Test Set 1: <u>higher IAA scores for document-level scenario for both binary and error type</u>
- Test Set 2: higher *K* for **error type** and higher IRR for sentence level. Doc-level higher K and IRR in binary



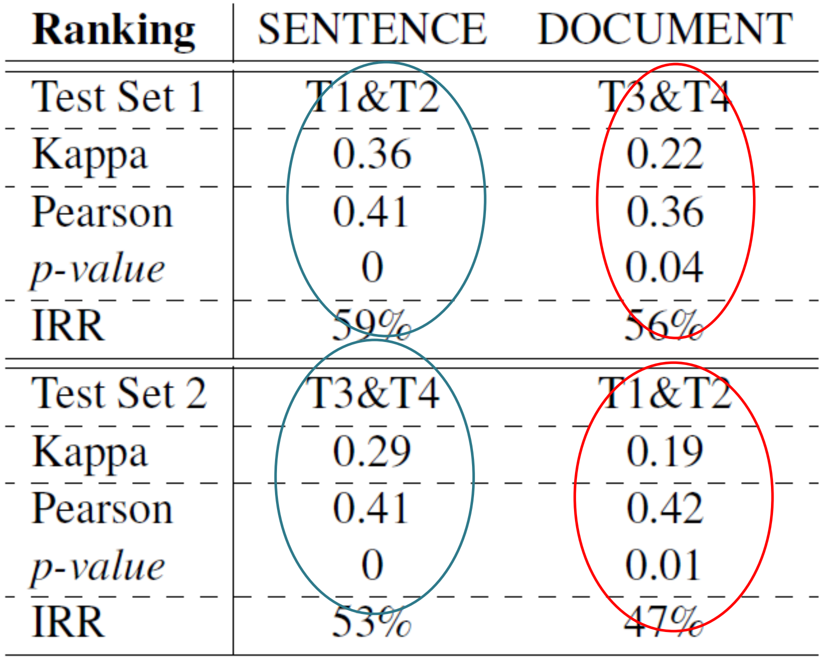
Error		SENTENCE	DOCUMENT
Test Set 1		T1&T2&T5	T3&T4
Kappa	binary	0.16	n/a
	type	0.02	0.31
IRR	binary	60%	100%
	type	56%	53%
Test Set 2		T3&T4	T1&T2&T5
Kappa	binary	0.49	-0.07
	type	0.38	-0.02
IRR	binary	76%	88%
	type	56%	50%

- T5: Test Set 2, IAA scores decrease for Kappa, both for binary and error type categories.
- IRR for binary slightly decreases

error markup at a document-level is difficult: tag problematic parts

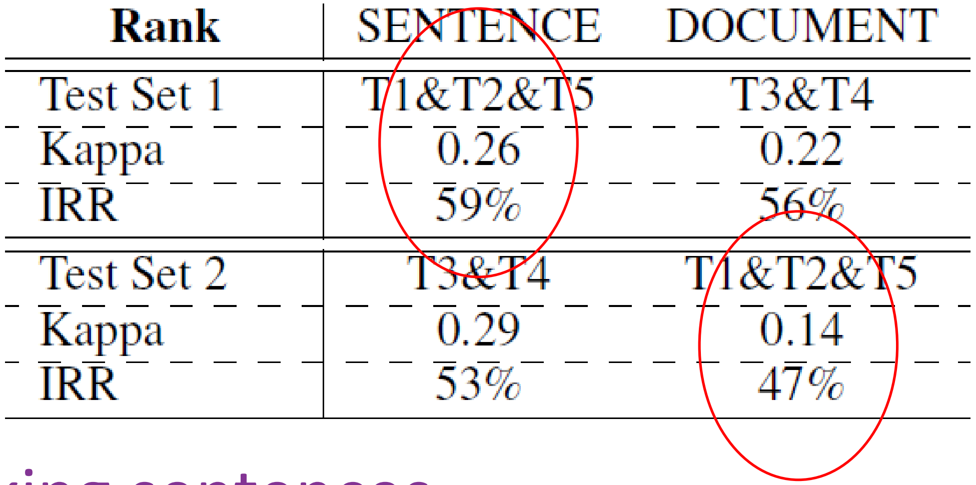
Results - IAA

- Ranking



Ranking	SENTENCE	DOCUMENT
Test Set 1	T1&T2	T3&T4
Kappa	0.36	0.22
Pearson	0.41	0.36
p-value	0	0.04
IRR	59%	56%
Test Set 2	T3&T4	T1&T2
Kappa	0.29	0.19
Pearson	0.41	0.42
p-value	0	0.01
IRR	53%	47%

- Test Set 1: higher IAA scores for sentence-level scenario
- Test Set 2: higher IAA scores for sentence-level scenario



Rank	SENTENCE	DOCUMENT
Test Set 1	T1&T2&T5	T3&T4
Kappa	0.26	0.22
IRR	59%	56%
Test Set 2	T3&T4	T1&T2&T5
Kappa	0.29	0.14
IRR	53%	47%

- T5: IRR scores do not change
- Slight decrease in *K* for both scenarios

IAA is higher when ranking sentences

Results - IAA

- Google seem to prefer to drop gender markers more than DeepL

1)**Source**: ***Her*** decision to pull out left everyone involved absolutely stunned.

DeepL: A decisão ***dela*** de se retirar deixou todos os envolvidos absolutamente atordoados.

Google: ***Sua*** decisão de sair deixou todos os envolvidos absolutamente atordoados.

2)**Source**: To recover ***it*** is a duty.”

DeepL: Recuperá-***lo*** é um dever”.

Google: Recuperar **(x)** é um dever.”

- Translators’ personal preferences
  - 1) adequacy over fluency
  - 2) drop of the gender marker when there is not enough context to specify the gender or solve ambiguity

Disagreements

- Disagreements in doc-level (especially for adequacy):
  - texts are made up of “very good”, “reasonably” and “poorly” translated sentences which, together, make the text understandable to a certain level, it is harder for translators to be consistent when assigning one single score for a full text.

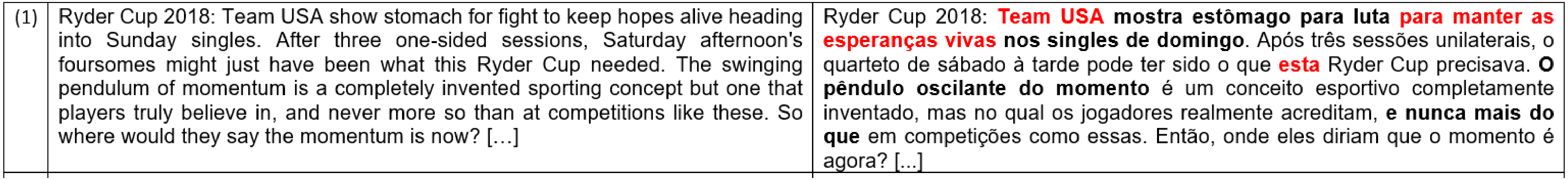
- Percentages of adequacy scores for the document-level scenario

How much of the meaning is in the translation?	T1&T2&T5	T3&T4
1 None of it	0%	4.69%
2 Little of it	7.29%	17.19%
3 Most of it	61.46%	64.06%
4 All of it	31.25%	14.06%

A great number of scores falling into the middle category makes it difficult for a consistent evaluation on a document-level scenario

Disagreements – document level

- Disagreements in doc-level (especially for adequacy)



(1)	Ryder Cup 2018: Team USA show stomach for fight to keep hopes alive heading into Sunday singles. After three one-sided sessions, Saturday afternoon's foursomes might just have been what this Ryder Cup needed. The swinging pendulum of momentum is a completely invented sporting concept but one that players truly believe in, and never more so than at competitions like these. So where would they say the momentum is now? […]	Ryder Cup 2018: **Team USA mostra estômago para luta para manter as esperanças vivas nos singles de domingo**. Após três sessões unilaterais, o quarteto de sábado à tarde pode ter sido o que **esta** Ryder Cup precisava. **O pêndulo oscilante do momento** é um conceito esportivo completamente inventado, mas no qual os jogadores realmente acreditam, **e nunca mais do que** em competições como essas. Então, onde eles diriam que o momento é agora? [...]

- Adequacy:
  - T1 : little of it - *“many mistranslations of golf/sport terms impaired meaning. Some unstranslated terms found (‘team USA’, ‘singles’)”*
  - T2: all of it - *“minor issues but the meaning isn’t lost”*
  - T5: most of it - *“the meaning is compromised by the word-by-word translation”*

- Fluency:
  - T1&T2: Near native
  - T5: Little Fluency – “Fluency compromised by the literal translation of some terms”

Disagreements – sentence level

- Disagreements in the sentence level are more often related to ambiguity and lack of context

(3) He then fired a beautiful through ball, leading Hazard into the box.	Ele então disparou uma bela bola cruzada, levando Hazard para dentro da caixa.

(3) Adequacy:

- T1 : none of it - *the translation "failed to use football terminology"*
- T2&T5: all of it

T2&T5 unaware that the sentence was about football due to the lack of context?

Mistranslations: 'fired' 'chutar' (to kick) and 'box' 'pequena área' not tagged

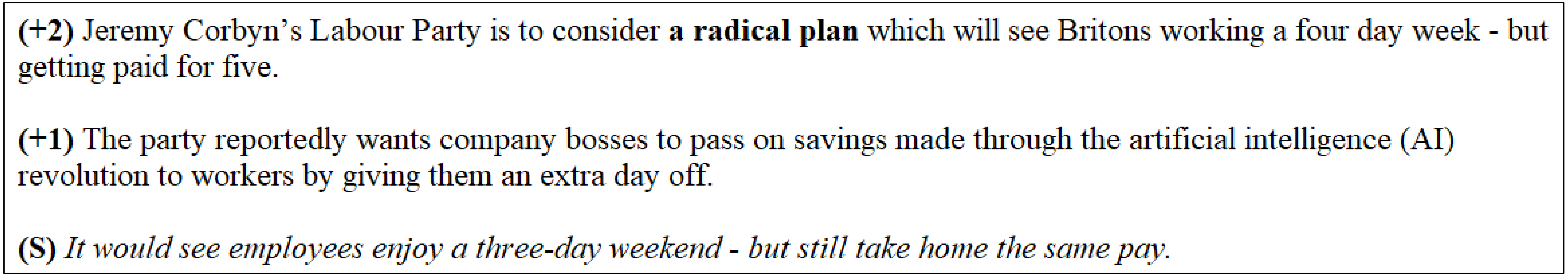
Out of context sentences leads to **misevaluation**!

Disagreements – sentence level

(4) It would see employees enjoy a three-day weekend – but still take home the same pay.	Veria que os funcionários desfrutariam de um fim de semana de três dias – mas ainda levariam para casa o mesmo salário.

Adequacy & Fluency

- T3: all of it & native
- T4: little & little - *“the context was not enough to translate the pronoun ‘it’”* (Castilho et al. (2020)



(+2) Jeremy Corbyn’s Labour Party is to consider **a radical plan** which will see Britons working a four day week - but getting paid for five.

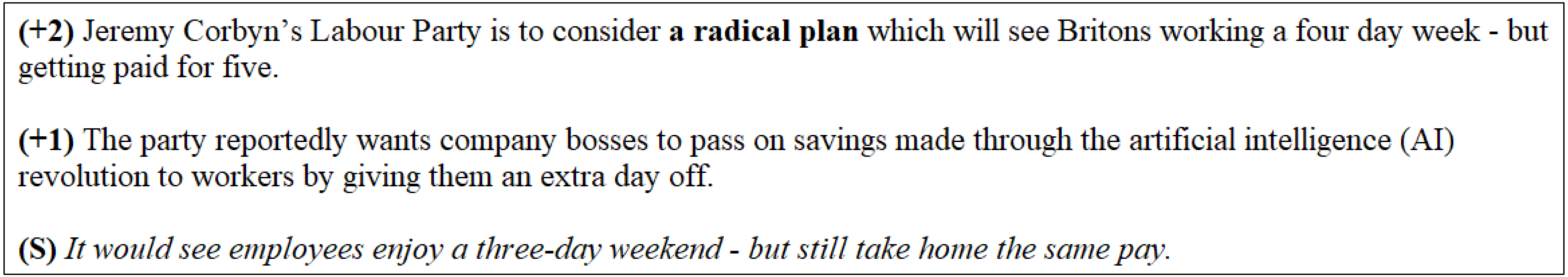
(+1) The party reportedly wants company bosses to pass on savings made through the artificial intelligence (AI) revolution to workers by giving them an extra day off.

(S) *It would see employees enjoy a three-day weekend - but still take home the same pay.*

T1, T2 and T5 assessment at doc-level: “near native” and contained “most” and “all” of the meaning.

Translators need access to a full context in order to assess the translation correctly

Disagreements – sentence level



(+2) Jeremy Corbyn's Labour Party is to consider **a radical plan** which will see Britons working a four day week - but getting paid for five.

(+1) The party reportedly wants company bosses to pass on savings made through the artificial intelligence (AI) revolution to workers by giving them an extra day off.

(S) *It would see employees enjoy a three-day weekend* - *but still take home the same pay.*

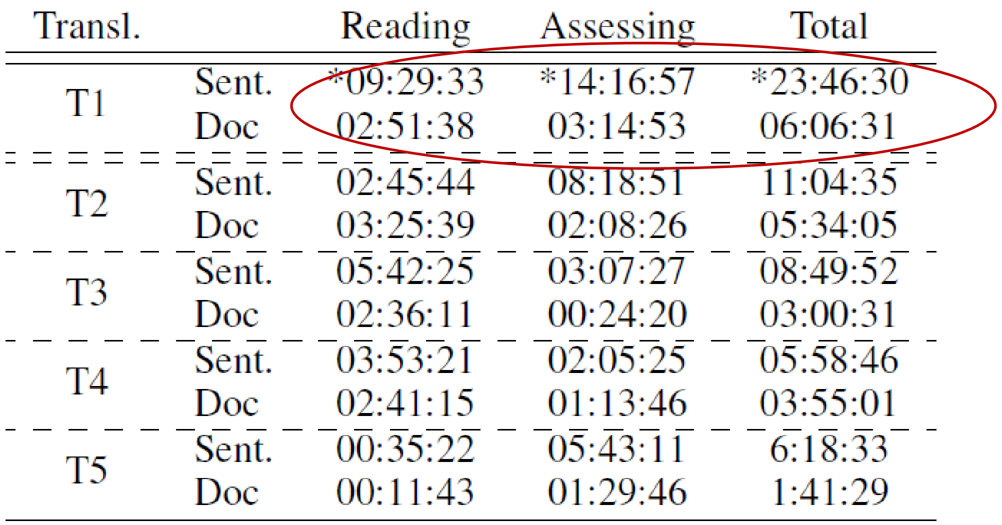
With the context of 2 previous sentences it is possible to identify that "it" relates to "a radical plan" and therefore the addition of "O plano veria" (the plan would see) in the translation would make it more adequate

- T1, T2 and T5 assessment at doc-level: "near native" and contained "most" and "all" of the meaning.

Translators need access to a full context in order to assess the translation correctly

Results - Effort

- Time – Not very conclusive



Transl.		Reading	Assessing	Total
T1	Sent.	*09:29:33	*14:16:57	*23:46:30
	Doc	02:51:38	03:14:53	06:06:31
T2	Sent.	02:45:44	08:18:51	11:04:35
	Doc	03:25:39	02:08:26	05:34:05
T3	Sent.	05:42:25	03:07:27	08:49:52
	Doc	02:36:11	00:24:20	03:00:31
T4	Sent.	03:53:21	02:05:25	05:58:46
	Doc	02:41:15	01:13:46	03:55:01
T5	Sent.	00:35:22	05:43:11	6:18:33
	Doc	00:11:43	01:29:46	1:41:29

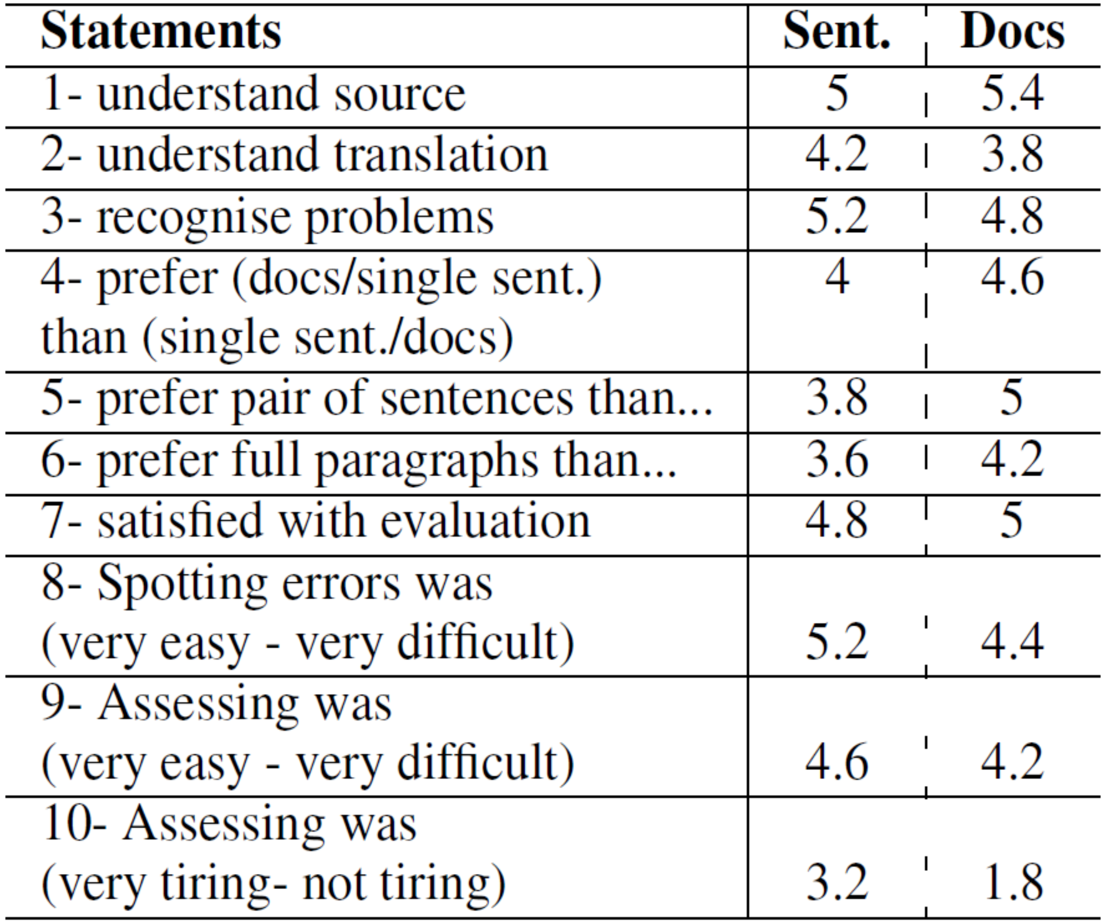
- Problems with PET UI
- Problems with T1

- Expectation:
  - longer reading time for the document-level scenario since full texts are longer
  - Longer assessing time for the sentence-level scenario

- Document-level scenario:
  - T1, T4 and T5 = lower reading time
  - T2 and T3 = longer reading time

Results - Effort

- Post-task Questionnaire



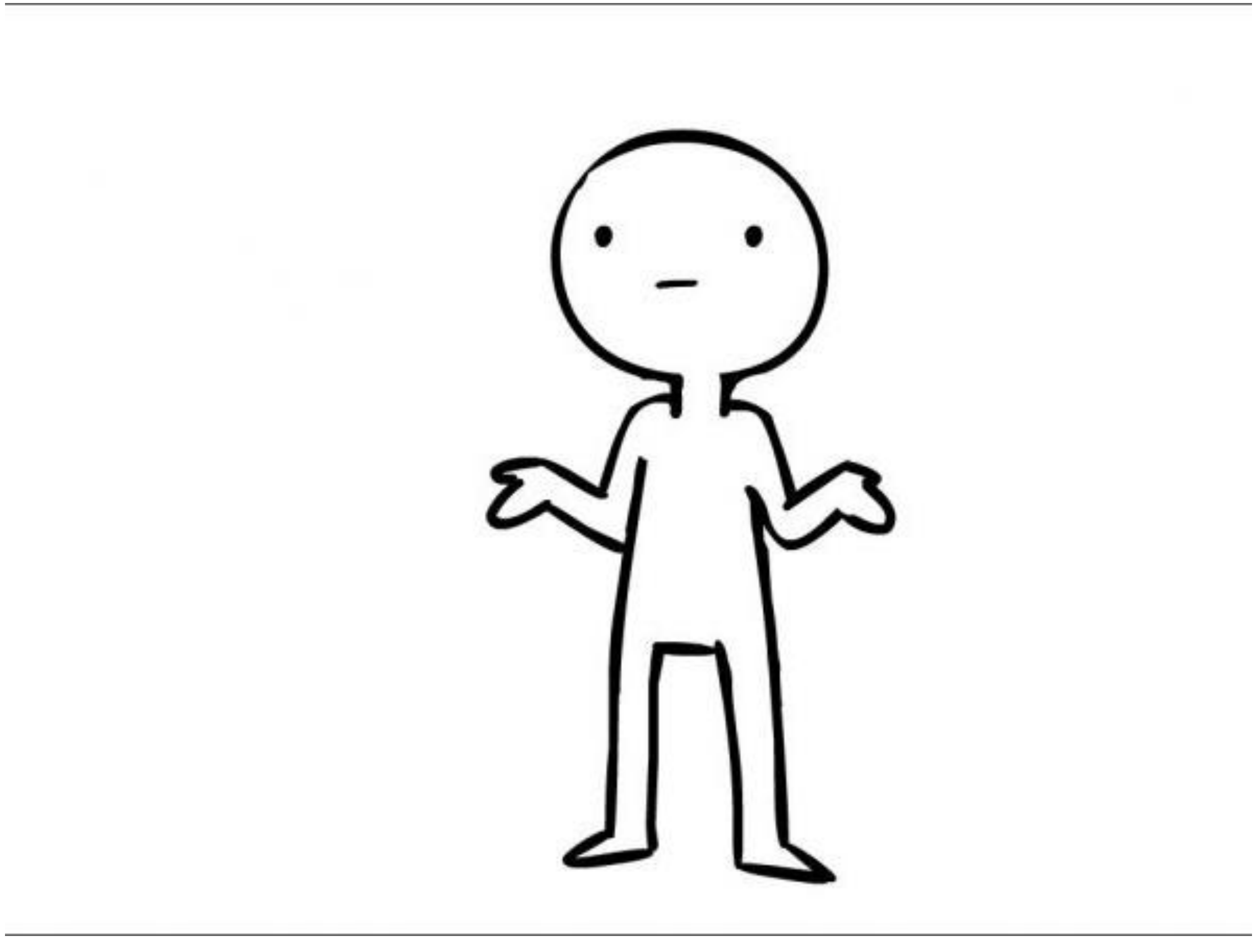
Statements	Sent.	Docs
1- understand source	5	5.4
2- understand translation	4.2	3.8
3- recognise problems	5.2	4.8
4- prefer (docs/single sent.) than (single sent./docs)	4	4.6
5- prefer pair of sentences than...	3.8	5
6- prefer full paragraphs than...	3.6	4.2
7- satisfied with evaluation	4.8	5
8- Spotting errors was (very easy - very difficult)	5.2	4.4
9- Assessing was (very easy - very difficult)	4.6	4.2
10- Assessing was (very tiring- not tiring)	3.2	1.8

Scale range from 1 to 6 where 1 is strongly disagree/very difficult/ very tiring and 6 is strongly agree/very easy/not tiring at all.

Translators seem to prefer/find:

- to judge single sentences than full documents (st. 4)
- sentence pairs (st. 5) or paragraphs (st. 6) than full documents.
- Easier to spot errors in the full texts (st.8) contradicting results for st. 3
- Document-level scenario to be slightly easier to assess (st. 9)
- Document-level scenario to be much more tiring than assessing single sentences (st. 10)

So, what is the best set-up for document-level Eval then?



What is the best set-up for document-level Eval?

- How much context do we need in a doc-level evaluation? Do we always need to show full documents? What about when it’s not possible? Will it vary according to domain? Language?

- one score per document – low IAA.
  - one score per sentence?
  - one score per chunks of texts?

- Bias – how can we balance?
  - Personal preferences
  - Lack of contexts makes evaluators accept MT output more

- Can one specific error in one sentence trigger bad scores for the whole document?
- Can one very good sentence trigger good scores for the whole document?
- Tiredness on whole text scores - yes
  - Cognitive load

What we found?

- A document-level evaluation methodology where translators assign one score per text leads to lower levels of IAA for adequacy, ranking, and error mark-up when compared to methodologies where translators assign one score per sentence

However…

- Without context to disambiguate some issues (gender, pronoun, ambiguity) translators might just trust the MT or go with their personal preferences
  - Leads to misevaluation

Final remarks

- Unsupervised human evaluations of MT is hard

- Lack of proper tool able to handle different MT evaluation methodologies makes the assessment even more complex

- As Google Translate seems to operate on a sentence-level, a document-level evaluation of adequacy is penalized since a document can be constituted of sentences with different levels of quality.

- Multiple scores per document might yield higher levels of IAA when compared to the randomized sentence-level set-up for both sentence and document-levels MT systems

tack
mulţumesc
dhanyavād
takk
teşekkür ederim
mahalo
Баярлалаа
hvala
감사합니다
dankie
gracias
thank you
Дякую
arigatô
dziękuję
ευχαριστώ
diolch
dank
tänan
спасибо
merci
köszönöm
kasih
Obrigada
谢谢
תודה
choukrane
shukran
a dank
děkuji
Շնորհակալություն
terima
kütos
Благодарам
grazie
gràcies
ačiū
Xièxiè
danke
ありがとう
tak
Mh'gōi
благодаря
ngiyabonga
paldies
grazzi
go raibh maith agat
kop khun
faleminderit
Ďakujem
хвала
nandri