Agree to Disagree: Analysis of Inter-Annotator Disagreements in Human Evaluation of Machine Translation Output

Maja Popović
ADAPT Centre
School of computing
Dublin City University, Ireland
maja.popovic@adaptcentre.ie

Abstract

This work describes an analysis of inter-annotator disagreements in human evaluation of machine translation output. The errors in the analysed texts were marked by multiple annotators under guidance of different quality criteria: adequacy, comprehension, and an unspecified generic mixture of adequacy and fluency. Our results show that different criteria result in different disagreements, and indicate that a clear definition of quality criterion can improve the inter-annotator agreement.

Furthermore, our results show that for certain linguistic phenomena which are not limited to one or two words (such as word ambiguity or gender) but span over several words or even entire phrases (such as negation or relative clause), disagreements do not necessarily represent "errors" or "noise" but are rather inherent to the evaluation process. On the other hand, for some other phenomena (such as omission or verb forms) agreement can be easily improved by providing more precise and detailed instructions to the evaluators.

1 Introduction

Despite the large number of automatic evaluation metrics designed for machine translation (MT) evaluation[1] which represent invaluable tools for rapid development of MT systems, human assessment of translation quality remains the gold standard, both for evaluating MT systems as well as for developing suitable automatic metrics. Human evaluation is typically provided in one of the following ways: assigning an overall quality score to each translated sentence, ranking two or more translations of the same source language sentence from best to worst, or annotating actual translation errors. The errors can be only highlighted (marked) or corrected (post-edited), but can also be classified according to a given pre-defined scheme, such as MQM scheme.[2]

However, human judgments of translation quality show a high degree of variance (Callison-Burch et al., 2008; Denkowski and Lavie, 2010), especially for fine-grained error classification based on a detailed error scheme involving many error types (Lommel et al., 2014; Klubička et al., 2018). One of the reasons for the variance is the variety of possible solutions: there is no single objectively correct translation of a given text, but rather a range of possible translations from perfect over good to acceptable. Moreover, there is no single universal criterion for translation quality. Although all manual evaluations are essentially based on some of the following three criteria: adequacy (meaning preservation; the most frequently used), fluency (grammar of the target language; frequently used) and comprehension (readability; rarely used), the precise definition of the criterion is not always given to the annotators. Very often, an unspecified mixture of adequacy and fluency is used. In addition, other factors like target audience, goal of translation, etc. can have influence on evaluator's perception of quality.

Most publications dealing with human evaluation of MT, such as (Vilar et al., 2007; Callison-Burch et al., 2008; Klubička et al., 2018; Kreutzer et al., 2020; Popović, 2020; Castilho, 2020; Freitag et al., 2021), report an overall inter-annotator agreement (IAA) score such as percentage of equal labels, Koehn's κ, Fleiss' κ, Krippendorf's α, or similar. However, less work has been done on analysing the actual disagreements. Ranking and assigning overall scores are the mostly used methods and a large amount of annotated data is publicly available (for example in WMT shared tasks[3]), but these methods do not provide enough information

[1] http://www.statmt.org/wmt20/metrics-task.html

[2] http://www.qt21.eu/mqm-definition/definition-2015-12-30.html

[3] http://www.statmt.org/wmt20/index.html

Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL), pages 234–243

for further analysis. On the other hand, error annotation (with or without classification) does. Some publications about error classification calculated separated IAA scores for different error categories and reported that some categories seem to be "more difficult" to agree on than others (Lommel et al., 2014; Klubička et al., 2018). However, to the best of our knowledge, systematic analysis of the relations between different linguistic phenomena and inter-annotator disagreements has not been carried out yet.

In this work, we analyse inter-annotator disagreement for a set of different linguistic phenomena. We used two publicly available data sets containing annotated errors without classification. We choose this type of error annotation because we believe that it can better reflect the differences in annotators' perception of errors than error classification: it is not bound to any predefined error scheme which might be tailored for a specific task and/or language pair, so that the evaluators had freedom to mark any part of the text which they perceive as problematic.

2 Related work

While publications usually report only an overall inter-annotator agreement, many researchers analysed the disagreements in natural language processing evaluation from different points of view.

Amidei et al. (2019) argues that standard IAA coefficients should not represent the only criterion for checking the reliability of human evaluation of natural language generation due to natural variability of human language, and suggest that correlation coefficients should be used. Some researchers compared IAA for different evaluation methods, such as ranking vs assigning overall scores. Belz and Kow (2010) compared the two methods for evaluating natural language generation, and Callison-Burch et al. (2008); Denkowski and Lavie (2010) for machine translation. All reported that evaluators generally agree more in ranking.

Castilho (2020) compares IAA for three evaluation methods for machine translation, namely ranking, assigning scores and error classification, but only in the context of evaluating isolated sentences vs evaluating larger amounts of text (paragraphs, "documents").

Recent studies which investigated disagreements in natural language inference (Pavlick and Kwiatkowski, 2019) and semantic annotation (Sommerauer et al., 2020) claim that disagreement in natural language evaluation is often expected due to ambiguity and variation of language. Therefore, a number of disagreements do not represent "errors" or "noise" but are fully legitimate. On the other hand, (Oortwijn et al., 2021) argue that inter-rater disagreements are not necessarily due to inherent properties of the language, but at least in part to the annotation task being underspecified.

However, none of the publications analysed the actual disagreements and error perception of different annotators. One of the first publications in this direction deals with IAA for error classification for MT using the MQM error scheme (Lommel et al., 2014). They reported that different degree of agreement can be observed for different error types, not always for the same reasons. For example, disagreement is high for word oder errors due to observed error span: while annotators agree that there are problems with word order in the given part of the sentence, they do not agree about exact words which should be marked as errors. On the other hand, evaluators often mixed up "mistranslation" and "terminology" because some evaluators found it difficult to distinguish the difference between the error types. The same error scheme was used in (Klubička et al., 2018). Although the focus was not on analysis of disagreements, Kappa coefficients were presented for different error types showing that the annotators agreed to a large extent on untranslated words as well as on number, gender or case errors, while most disagreements were coming from omissions and tense errors. Contrary to (Lommel et al., 2014), the agreement for order errors was high, however this analysis was carried out on the sentence level, not on the word level, therefore diminishing the word-level span problem.

Another publication dealing with different error types reports results for natural language generation (Thomson and Reiter, 2020). Similarly to (Lommel et al., 2014), they report that some error types are more difficult for annotators to agree on than others, as well as the word span plays an important role for disagreements. Also they say that some of the observed disagreements could have been resolved by more detailed annotation instructions, while others are more fundamental.

To the best of our knowledge, none of the publications analysed inter-annotator disagreements in terms of underlying linguistic phenomena.

3 Data sets

The main part of our experiments was carried out on the "*QRev*" data set (Popović, 2020).[4] The set consists of English user reviews translated into Croatian and Serbian. For each of the target languages, five different NMT systems were used: three online systems (*Amazon Translate*, *Microsoft Bing* and *Google Translate*) and two in-house systems (Popović et al., 2021) based on the Sockeye[5] implementation, one designed for translating general domain, and the other for translating user reviews. In total, the data set contains outputs of ten different MT systems.

Two quality criteria were used for highlighting errors: adequacy and comprehension. An important difference between the two criteria is that seeing the source text was *required* for marking adequacy errors while seeing the source text was *forbidden* for marking comprehension errors. For both quality aspects, the evaluators were asked to concentrate on problematic parts of the text and to highlight them. For adequacy, they were instructed to highlight parts which entirely or partially change the meaning of the source text. For comprehension, they were asked to mark parts which are impossible or hard to understand. All translations were evaluated in context – the evaluators were seeing entire reviews.

In total, 15 evaluators, computational linguistics students and researchers fluent in the source language and native speakers of the target language, participated in the annotation. The largest part of the text (about 3000 sentences) is annotated by two evaluators, while a small part of the text (about 40 sentences) is annotated by three or four evaluators. Inter-annotator agreement in terms of Krippendorf's α is 0.61 for adequacy errors and 0.51 for comprehension errors.

We also worked on a small "*HumanMT*"[6] (Kreutzer et al., 2020) data set consisting of English TED talks translated into German by one in-house MT system. This data set was not created for purposes of MT evaluation, but for improving an NMT system by giving it feedback about errors. Since the used loss function did not support omissions and reordering errors, the evaluators are specifically asked not to highlight these two types of errors. The evaluators were not given any specific quality criterion: they were only instructed to "highlight the errors", which usually implies a mixture of adequacy and fluency in MT evaluations, but was not specified. The part of this corpus annotated by multiple evaluators is very small, only 11 sentences, however each of them was annotated by 5-10 evaluators, university students with fluent or native English and German skills. The reported Krippendorf's α is much lower, 0.201.

Due to all above mentioned limitations of the corpus, it is not sufficient to draw any conclusions. Still, qualitative inspection of inter-annotator disagreements was helpful to confirm some of the hypotheses based on the results on the *QRev* corpus.

An overview of the two data sets is presented in Table 1.

data set	*QRev*	*HumanMT*
language pair	en→sr,hr	en→de
domain	user reviews	TED talks
# of MT systems	10	1
# of unique segments	3334	11
total # of annotators	15	10
# of annotators per segment	2	5-10
quality criterion	adequacy, comprehension	not specified
Krippendorf's α	adeq. 0.610 compr. 0.510	0.201

Table 1: Statistics of the two analysed data sets.

4 Analysis of disagreements

In order to analyse the nature of differences in error perception between two evaluators, the first step was to identify the nature and causes of the highlighted errors, namely the underlying linguistic phenomena. The second step was then to calculate the overlap of words perceived as errors by two annotators for each identified underlying phenomenon; higher overlap indicates higher agreement between the annotators.

4.1 Underlying linguistic phenomena

We analysed the nature of the highlighted errors by tagging them with possible causes and/or plausible explanations of their origin (referred to as "underlying linguistic phenomena" or simply "phenomena"). The definition of these phenomena is

[4] https://github.com/m-popovic/QRev-annotations

[5] https://github.com/awslabs/sockeye

[6] https://www.cl.uni-heidelberg.de/statnlpgroup/humanmt/

based both on general linguistic knowledge as well as on phenomena related to the (machine) translation process. We did not have any pre-defined scheme for the phenomena, but we started by looking into errors and identifying the phenomena on the fly.

The errors in the described data sets were analysed in the following way: they were tagged as a particular phenomenon if 1) they were marked by at least one evaluator 2) it was possible to define a plausible cause and/or explanation for their origin.

In total, we identified 26 phenomena. Some of them are equivalent to the typical error types in error classification tasks, such as “mistranslation", “case", “word order", “ambiguous word". These usually involve only a single word or a small group of two or three words. Others are going far beyond typical error classification, often bringing on several different interwining types of errors. These phenomena encompass larger spans of words, sometimes even entire sentences. However, not necessarily all the involved words are perceived as errors. Since the evaluation protocol allowed free annotation of errors, evaluators often highlighted different words in order to mark the same issue, thus agreeing about the existence of the problem but disagreeing on the precise locations and span. Sometimes, evaluators marked many consecutive words although only a few of them were actually erroneous, which we refer to as“error propagation". These “propagated" errors could not be interpreted by any linguistic phenomenon and are tagged as “None". Error propagation was found much more frequently in comprehension errors. Another type of errors with the “None" tag are related to the individual stylistic preferences of different annotators, and are equally frequent for both quality criteria. In total, about 25% of all words marked as adequacy errors and 38% of comprehension errors belong to this category.

While the majority of phenomena is self explained by their name, we explain the more complex large span phenomena in details. Table 2 illustrates four large span phenomena: rephrasing, noun phrases, conjunction (before relative clause) and negation. Errors related to the particular phenomenon are marked in bold, and omissions are marked as “X".

Rephrasing (Table 2(a)) refers to a sequence of source words which is not translated properly for some of the following reasons or their combination: (i) rephrasing is needed in the target language but the translation follows the structure of the source language (examples 1 and 3) (ii) rephrasing is not needed in the target language but is applied (iii) rephrasing is needed in the target language but it is incorrectly applied (example 2). The phenomenon also comprises incorrect translation of multi-word expressions.

Noun phrase is a similar phenomenon but it encompasses small number of words and refers particularly to a head noun together with additional nouns and adjectives. Examples can be seen in Table 2(b).

Table 2(c) presents issues related to *conjunction* preceding a relative clause. If conjunction before a relative clause in the source language is omitted (which happens frequently in English), it can result in incorrect translation of several words around the conjunction (example 1), especially if the target language requires a conjunction (which is the case for all three analysed languages). And sometimes the reverse problem occurs: there is no relative clause in the source sentence, but a spurious conjunction appears in the translation (example 2).

Negation is another phenomenon with a large span often involving an entire sentence. Sometimes only a single word or two are incorrectly translated (example 1), but sometimes the entire sentence (example 2). Sometimes just a negation mark is missing (example 3) or inserted.

4.2 Overlap of words marked by two evaluators

In order to estimate the overlap of errors perceived by two different annotators, the following formula was used:

$$overlap = \frac{2 * C(words_{ev1+ev2})}{C(words_{ev1}) + C(words_{ev2})}$$

where $C(words_{ev1+ev2})$ denotes the number (count) of words which are perceived as errors by both evaluators, $C(words_{ev1})$ is the count of words perceived as errors by the first evaluator and $C(words_{ev2})$ the count of words perceived as errors by the second evaluator. Examples of words marked by two evaluators and their overlap for ten sentences/segments can be seen in Table 3. Only the marked words were extracted from the text, and the words marked by both evaluators are presented in bold.

(a) rephrasing

language	group of words to be rephrased
1) EN source	it does a good job of protecting
SR/HR correct	dobro štiti
MT outputs	**to radi dobar posao štititi**
2) EN source	gets his little gray cells working
SR/HR correct	aktivira svoje male sive ćelije
MT output	**radi na svojim malim sivim ćelijama**
EN gloss	works on his little gray cells
3) EN source	you name it
DE correct	was (auch immer) Sie wollen
MT output	**Sie bennenen es**

(b) noun phrases

language	noun phrase
1) EN source	bird feeder
SR/HR correct	hranilica za ptice
MT outputs	hranilica **X ptica**
	ptica hranilica
2) EN source	traveling salesman problem
DE correct	Problem des Handlungsreisenden
MT output	**Reisen Verkäufer** Problem

(c) conjunction

language	relevant parts of the sentence
1) EN source	For the kind of shipping they want it would be reasonable to expect
SR/HR correct	Za vrstu dostave koju žele bilo bi razumno očekivati
MT output	Za vrstu dostave **žele da bi bilo** razumno očekivati
2) EN source	The other DVDs he doesn't even look at
SR/HR correct	Ostale DVDove on ni ne pogleda
MT outputs	**Ostali DVDovi koje** on ni ne gleda
	Ostale DVD-ove **koje** on ni ne gleda
EN gloss	The other DVDs **which** he doesn't look at

(d) negation

language	negation span
1) EN source	I never liked any of the
SR/HR correct	Nikad nisam volio niti jedan
MT output	Nikad nisam volio **bilo koji**
2) EN source	without me even drinking anything
SR/HR correct	čak iako nisam ništa pio
MT outputs	**bez mene čak i pio ništa**
	bez mene ni da ništa ne pijem
3) EN source	women are misdiagnosed
DE correct	Frauen werden Fehldiagnosen gestellt
MT output	Frauen werden **X** diagnostiziert

Table 2: Examples of phenomena involving several consecutive words and different types of errors: noun phrases (a), rephrasing (b), conjunction (c) and negation (d). Errors related to the given phenomenon are marked in bold; missing parts are denoted by "X".

	words marked by evaluator #1	words marked by evaluator #2	overlap
1)	**ovaj**	**ovaj**	100
2)	**na brigu koju sam** **koristio želio je** naginjati	**na brigu koju sam** **koristio** u bacanju **želio je**	82.4
3)	**Amazone** nađete	**Amazone**	66.7
4)	ovdje je **je**	X **je**	40.0
5)	To je **širok**	komada uklopiti **širok**	33.3
6)	zombi film		0
7)	**so bin ich I**	kanalisiert Prozess verändert **so bin ich I**	72.7
8)	**wieder**	**wieder** zu	66.7
9)	spiele **Übung**	**Übung** üben	50.0
10)	**String-Instrument** zu einer Zeit	begrenzen aus **String-Instrument** kann	25.0

Table 3: Examples of words marked by two different evaluators and the calculated overlap; overlapping words are presented in bold; "X" stands for omission.

In sentence 1), both evaluators marked the same word "ovaj", so that the overlap is 100%. In sentence 2), 7 words are highighted by both evaluators, while one word ("naginjati") is marked only by the evaluator #1 and two words ("u", "bacanju") only by the evaluator #2, exhibiting a high overlap of 82.4%. In sentence 6), the first evaluator marked two words while the second one did not mark anything, therefore the overlap is 0. The rest of sentences shows different levels of overlaps between 20 and 70%.

5 Results

For each of the two quality criteria, the overall overlap (agreement) together with the overlap for the most interesting identified phenomena for the *QRev* data set are presented in Table 4. A phenomenon is considered as interesting if (i) the overlap is generally low or high (ii) the overlap is different for the two quality criteria (iii) the phenomenon appears frequently in the data. We did not analyse only frequent phenomena, because the frequency of errors is not necessarily related to their importance or severity (Federico et al., 2014; Kirchhoff et al.,

	adequacy	comprehesion
overall	59.3	56.9
omission$*$	34.7	20.1
conjunction^{--}	59.7	66.7
order^{--}	60.4	63.9
negation^{-}	63.4	65.3
tense/aspect/mood^{-}	64.9	69.7
named entity*	67.8	68.1
rephrasing**	68.4	69.8
noun phrase**	72.0	68.4
ambiguity**	75.2	72.3
gender*	76.2	72.8
case*	77.9	81.7
person $^{-}$	83.1	77.0
untranslated*	83.5	84.3
mistranslation*	83.6	77.8
-ing^{-}	83.6	84.9
source error^{-}	83.8	80.2
non existing word^{-}	84.9	88.9
hallucination^{--}	85.7	0
none	19.7	28.2

Table 4: Overlap (%) of words marked as adequacy (left) and comprehension (right) errors by two different evaluators in the *QRev* data set: overall and for different phenomena ordered from lowest to highest.

2014). It should be, however, taken into account that for the less frequent phenomena the results of the presented analysis might be less reliable. For these reasons, we marked the frequency of each phenomenon in the following way: "$**$" denotes phenomena which contrubute to more than 5% of all highlighted words, "$*$" refers to those contributing to 2-5% of words, "$-$" to 1-2% of words, and "$--$" to less than 1%.

5.1 Influence of quality criterion

First of all, it can be noted that there is a very low overlap for the tag "None", which could be expected since, as explained in Section 4.1, these annotations are related to evaluators' stylistic preferences as well as different perceptions of the word span.

Since adequacy is the most widely used criterion in machine translation evaluation, the phenomena in Table 4 are ordered from the highest to the lowest disagreement for adequacy error marking. The overall tendencies are similar for comprehension, but there are some differences. First, overall agreement is lower for comprehension, mainly due to the stronger effect of "error propagation" described in Section 4.1: because the evaluators do not see the original source text while annotating comprehension issues, there is more freedom and more room for subjectivity. Also, some of the phenomena have a notably different overlaps for the two phenomena than others which will be discussed in the next section.

5.2 Influence of underlying phenomena

Low agreement The lowest overlap (largest disagreement) can be noted for ***omissions***, especially for comprehension but also for adequacy. This is not surprising for comprehension, because the source text is not available so that it is not possible to see what is really omitted from the source text and what only looks like something is missing. As for adequacy, one of the reasons is that the evaluators were not specifically instructed to distinguish cases when content in the source text is missing from cases when something related only to the target text is missing (such as auxiliary verb or preposition). Furthermore, multiple omissions (several missing words) were marked differently: some evaluators inserted only one omission marker even for word groups/phrases, others inserted one omission marker for each of the missing words, while some inserted two markers to indicate a multiple omission irregardless of the actual number of missing words. These findings indicate that more precise instructions for marking omissions can increase the agreement. The omission example 1 in Table 5 illustrates the case where one evaluator inserted an omission mark for each of several consecutive missing words while the other inserted only one omission mark. In example 2, the first evaluator did not inserted any omission marks but highlighted the (correct) words surrounding the omissions, while the second one inserted two omission marks, one for each missing word.

Word order is another type of issues with low overlap which has already been known for differences in perception of the exact span and involved words. In the example 3 in Table 5, the first annotator marked only one word which can be moved to resolve the error, while the second annotator marked the entire phrase.

Perception of word span is especially different for all complex phenomena encompassing larger word spans and different types of errors described in Section 4.1: ***conjunction***, ***negation***, ***rephrasing***, and to a lesser extent (due to a shorter word span) ***noun phrases***. For all of them except noun phrases, agreement is lower for adequacy, and this difference is especially high for conjunction. These

adequacy disagreements are partly due to different ideas about the correct translation: looking at the source text, each evalutor has a different correct translation of it in mind and annotates errors according to it. Another reason is different perception of units in the text, which happens both for adequacy and for comprehension. All these disagreements are subjective to a large extent and cannot be completely avoided.

For *conjunction* related errors in the example 4, Table 5, the evaluators disagree about the number of erroneous words. In example 5, the evaluators agree only on two words. They both mark several surrounding words, but one annotator perceived the left part as problematic while the other one marked the right part of the phrase. Similar tendencies can be noted in the *negation* examples 6, 7 and 8.

The *rephrasing* example 9 demonstrates differences in span, word positions as well as perception of error types. The evaluators agree only on one word. The first evaluator marked several surrounding words while the second one inserted two omission markers. In the example 10 illustrating a *noun phrase*, one evaluator considered both words in the phrase as errors while another marked only one word.

Lower agreement can also be noted for ***named entities***, partly because they often consist of several words thus inducing span related disagreements (example 11), but evaluator personal preferences can be noted, too. For some named entities there is no standard in the target language, so one evaluator might prefer the original English name and another would rather see it translated. In example 12, two instances of a game name appeared in one sentence: once the full name (“Last Night on Earth") and once the abbreviation “LNOE". The full name was translated by MT, while the abbreviation was copied. Then, the first evaluator preferred the original so they marked the translated version as error, while the second evaluator preferred the translated name and marked the original as error.

Verb forms (***tense/aspect/mood***) also result in a number of disagreements, mostly related to span: in the example 13, the first evaluator marked the entire verb phrase while the second one marked only the incorrect form of the auxiliary verb. Furthermore, the disagreement regarding verb form errors is notably lower for adequacy. Qualitative analysis showed that if a tense in the translation is the same as the tense in the source text, evaluators sometimes do not perceive it as an error even when it is incorrect in the target language.

High agreement Several phenomena have very high agreement (over 80%), most of them involve only one or two words: ***untranslated*** words (copied from the source text into the translation), ***mistranslation***, error in the ***source text*** and ***non-existing word***. A high overlap over 70% can be noted for ***ambiguous words***, ***gender*** and ***case*** errors.

Finally, ***hallucination*** represents an interesting and specific phenomenon. It refers to the parts of the translated text which have no connection to the original source text. Since it deteriorates adequacy by the definition (because the meaning of the original text is changed), the annotators overly agree about it being an adequacy error. However, the situation is different for comprehension: one evaluator perceived these errros as fully comprehensible, while another one noted that something is not right and marked them as errors. The reported comprehensibility agreement of 0% should not, however, be taken as absolute truth since hallucinations are very rare in the analysed data. Still, it should be taken into account that without the access to the source text, some annotators might perceive them as fully comprehensible.

5.3 HumanMT data set

As already mentioned, we also analysed the small *HumanMT* data set, although it is not convenient for drawing any conclusions. Still, several interesting tendencies were observed which relate to the previously described findings as well as the previous work.

In this corpus, *untranslated* words also have a high overlap, while long span phenomena such as *rephrasing* and *noun phrases* have a low overlap. However, some of the short span phenomena such as *ambiguity* or *gender* also have a low overlap. Quality inspection of the annotations showed that the lack of a precisely defined quality criterion contributed to the generally low agreeement, because much more subjectivity and personal preferences are allowed for all phenomena. Not marking omissions and ordering errors also contributed to lower agreement because (i) some of the evaluators still marked order-related errors while others did not (ii) some annotators marked correct words around an omission (similarly to the example 2 in Table 5).

omission	1) source	if they can be actresses everybody can!
	evaluator #1	**X X X X X** svi mogu!
	evaluator #2	**X** svi mogu!
	2) source	What the heck is this show about?
	evaluator #1	O čemu **se radi ova emisija**?
	evaluator #2	O čemu se **X** radi **X** ova emisija?
order	3) source	it helped me
	evaluator #1	mi je **pomogao**
	evaluator #2	**mi je pomogao**
conjunction	4) source	The other DVD's he doesn't even look at.
	evaluator #1	Ostali **DVD-ovi koje** on ni ne gleda.
	evaluator #2	**Ostali DVD-ovi koje** on ni ne gleda.
	5) source	after a trip the girls take to Palm Springs
	evaluator1	nakon putovanja **djevojke idu u Palm Springs**
	evaluator2	**nakon putovanja djevojke idu** u Palm Springs
negation	6) source	Doesn't have the same feel either.
	evaluator #1	Nema isti osjećaj **bilo**.
	evaluator #2	**Nema isti osjećaj bilo**.
	7) source	I find nothing redeeming about any of these characters or care about anything they do
	evaluator #1	ne nalazim ništa otkupljujuće o bilo kojem od tih likova ili **stalo do bilo čega što rade**
	evaluator #2	ne nalazim ništa otkupljujuće o bilo kojem od tih likova **ili stalo do bilo čega** što rade
	8) source	they were both non-responsive
	evaluator #1	oboje **nisu reagirali**
	evaluator #2	**oboje nisu** reagirali
rephrasing	9) source	Most of the flavors taste nothing like their names
	evaluator #1	Većina okusa nema **ništa X** poput **X** imena
	evaluator #2	Većina okusa **nema ništa poput imena**
noun phrase	10) source	it is no better than the store products
	evaluator #1	nije ništa bolji od **prodajnih proizvoda**
	evaluator #2	nije ništa bolji od **prodajnih** proizvoda
named entity	11) source	apart from Austin Powers
	evaluator#1	osim **Austin** Powersa
	evaluator#2	osim **Austin Powersa**
	12) source	Last Night on Earth, LNOE
	evaluator #1	Sinoć na Zemlji, **LNOE**
	evaluator #2	**Sinoć na Zemlji**, LNOE
tense/aspect/mood	13) source	he could know
	evaluator #1	**mogao je da zna**
	evaluator #2	mogao **je** da zna

Table 5: Examples of disagreements between two evaluators regarding adequacy errors for phenomena with the lowest overlap: omission, word order, conjunction, negation, rephrasing, noun phrase, named entity, tense/aspect/mood. Marked errors are presented in bold; "X" stands for omission.

6 Conclusions

This work attempts to shed light on the differences between error perception of different annotators for evaluation of machine translation output. The main findings are that the quality criterion as well as underlying linguistic phenomena have a strong influence on error perception. For some of the phenomena, such as omission or verb forms, agreement can be increased by providing more detailed annotation instructions. For others, especially those with a large word span such as negation or rephrasing, the differences are inherently subjective and therefore hard to completely avoid. This is, nevertheless, not necessarily bad, but it is important to be aware of it. While improved agreement certainly is an important goal, the exact nature of disagreements where perception of errors does not result in the same annotation can also provide insight into how humans perceive the translation quality.

Based on these findings, our recommendations for human evaluation are:

1) Define a quality criterion and provide detailed description of it;

2) Pay attention to phenomena which can benefit from precise instructions, such as omission, named entities or verb forms (tense/aspect/mood);

The instructions for omissions could be:

- for each omitted word in the source language, insert the omission mark “X”;
- for each word missing in the target language, insert the omission mark “Y"

For named entities:

- mark each translated person name as error;
- mark each translated music band name as error;
- mark each untranslated movie title as error; etc.

For verb forms:

- mark each incorrect individual component of verb tense as an error; if all components are incorrect, mark all;
- if auxiliary or main verb are missing, insert omission mark “X";
- if a verb form does not seem natural in the target language, mark it as error even though it corresponds to the verb form in the source language; etc.

3) If possible, try to increase the agreement for complex long span phenomena, too, but do not worry about certain inavoidable amount of disagreements.

Acknowledgements

The ADAPT SFI Centre for Digital Media Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant 13/RC/2106. This research was partly funded by financial support of the European Association for Machine Translation (EAMT) under its programme “2019 Sponsorship of Activities”.

References

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2019. Agreement is overrated: A plea for correlation to assess human evaluation reliability. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 344–354, Tokyo, Japan.

Anja Belz and Eric Kow. 2010. Comparing rating scales and preference judgements in language evaluation. In *Proceedings of the 6th International Natural Language Generation Conference (INLG 10)*.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation (WMT 08)*, pages 70–106, Columbus, Ohio.

Sheila Castilho. 2020. On the same page? comparing inter-annotator agreement in sentence and document level human machine translation evaluation. In *Proceedings of the Fifth Conference on Machine Translation (WMT 20)*, pages 1150–1159, Online.

Michael Denkowski and Alon Lavie. 2010. Choosing the right evaluation for machine translation: an examination of annotator and automatic metric performance on human judgment tasks. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas (AMTA 10)*, Denver, Colorado, USA.

Marcello Federico, Matteo Negri, Luisa Bentivogli, and Marco Turchi. 2014. Assessing the impact of translation errors on machine translation quality with mixed-effects models. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 14)*, pages 1643–1653, Doha, Qatar.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. In *Arxiv*.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2018. The sockeye neural machine translation toolkit at AMTA 2018. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (AMTA 18)*, pages 200–207, Boston, MA.

Katrin Kirchhoff, Daniel Capurro, and Anne M. Turner. 2014. A conjoint analysis framework for evaluating user preferences in machine translation. *Machine Translation*, 28(1):1–17.

Filip Klubička, Antonio Toral, and Víctor M. Sánchez-Cartagena. 2018. Quantitative Fine-grained Human Evaluation of Machine Translation Systems: A Case Study on English to Croatian. *Machine Translation*, 32(3):195–215.

Julia Kreutzer, Nathaniel Berger, and Stefan Riezler. 2020. Correct Me If You Can: Learning from Error Corrections and Markings. *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT 20)*.

Arle Lommel, Maja Popović, and Aljoscha Burchardt. 2014. Assessing inter-annotator agreement for translation error annotation. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 14)*, Reykjavik, Iceland.

Yvette Oortwijn, Thijs Ossenkoppele, and Arianna Betti. 2021. Interrater disagreement resolution: A systematic procedure to reach consensus in annotation tasks. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval 21)*, pages 131–141, Online.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Maja Popović. 2020. Informative manual evaluation of machine translation output. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 20)*, Online.

Maja Popović, Alberto Poncelas, Marija Brkić Bakarić, and Andy Way. 2021. On machine translation of user reviews. In *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP 21)*, Online.

Pia Sommerauer, Antske Fokkens, and Piek Vossen. 2020. Would you describe a leopard as yellow? evaluating crowd-annotations with justified and informative disagreement. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 20)*, pages 4798–4809, Barcelona, Spain (Online).

Craig Thomson and Ehud Reiter. 2020. A gold standard methodology for evaluating accuracy in data-to-text systems. In *Proceedings of the 13th International Conference on Natural Language Generation (INLG 2020)*, pages 158–168, Dublin, Ireland.

David Vilar, Gregor Leusch, Hermann Ney, and Rafael E. Banchs. 2007. Human evaluation of machine translation through binary system comparisons. In *Proceedings of the Second Workshop on Statistical Machine Translation (WMT 07)*, pages 96–103, Prague, Czech Republic.