On the practice of error analysis for machine translation evaluation

Sara Stymne, Lars Ahrenberg

Linköping University
Linköping, Sweden
{sara.stymne,lars.ahrenberg}@liu.se

Abstract

Error analysis is a means to assess machine translation output in qualitative terms, which can be used as a basis for the generation of error profiles for different systems. As for other subjective approaches to evaluation it runs the risk of low inter-annotator agreement, but very often in papers applying error analysis to MT, this aspect is not even discussed. In this paper, we report results from a comparative evaluation of two systems where agreement initially was low, and discuss the different ways we used to improve it. We compared the effects of using more or less fine-grained taxonomies, and the possibility to restrict analysis to short sentences only. We report results on inter-annotator agreement before and after measures were taken, on error categories that are most likely to be confused, and on the possibility to establish error profiles also in the absence of a high inter-annotator agreement.

Keywords: Error analysis, Inter-annotator agreement, Machine translation evaluation

1. Introduction

Error analysis is the identification and classification of individual errors in a machine translated text. Such an evaluation can point to specific strengths and problem areas for a machine translation system, which is hard to do using standard automatic evaluation metrics like Bleu (Papineni et al., 2002) or human ranking of sentences (Callison-Burch et al., 2007). It thus provides a better foundation for decision-making, whether relating to system development, purchase, or use. By tabulating the distribution of errors that a system makes over the different error categories, we obtain an error profile.

Virtually all approaches to human evaluation of machine translation output suffer from the weakness of low inter-annotator reliability. However, in papers applying error taxonomies to machine translation output, we rarely find discussions of this issue, or reports of agreement metrics. In this paper we want to do just this, using data from a study aiming at profiling two versions of the same system. In particular we want to answer the following questions:

- What levels of inter-annotator reliability can be expected, and how do the levels depend on the properties of the error taxonomy?
- What measures can be taken to improve inter-annotator reliability?
- Can we distinguish two versions of the same system qualitatively using an error profile, and if so how?
- Can we draw any conclusions about the performance of the two systems, even if inter-annotator reliability is low, and, if so, how?
- Can we draw the same conclusions from a sample of short sentences, which are easier to annotate, as from a random sample of sentences?

2. Related work

There have been several suggestions of taxonomies for MT error analysis (Flanagan, 1994; Elliott et al., 2004; Vilar et al., 2006; Farrús et al., 2010). Especially the taxonomy by Vilar et al. (2006) has been used by several other researchers, e.g. Avramidis and Koehn (2008) and Popović and Burchardt (2011).

We have not been able to find any work that report inter-annotator agreement when using such taxonomies, however, though Elliott et al. (2004) mentions it as future work. Popović and Burchardt (2011) notes that "human error classification is definitely not unambiguous" and mentions some categories even at top level that tend to get confused. There are also several error taxonomies for classification of human translation errors, see Secară (2005) for a summary.

For many other aspects of MT evaluation, inter-annotator agreement has been discussed extensively. Callison-Burch et al. (2007) found that inter-annotator was low for several tasks; they reported kappa figures of .25 for fluency judgments, .23 for adequacy judgments and .37 for ranking of full sentences. When only considering ranking of high-lighted constituents, kappa went up to .54. They also found inconsistencies when the same annotator was presented the same sentences several times, the kappa scores for this ranged between .47–.76 for the tasks mentioned.

Another problem with human evaluation is that it takes time and human resources. Popović and Burchardt (2011) showed that automatic error classification measures can be developed that correlate well with human classification. They investigated two types of human annotations, one that strictly compared a system translation with a reference sentence, and one flexible annotation type where differences from the reference sentences were allowed if they were syntactically and semantically correct. The automatic error classification had both a high precision and recall on the strict annotations, but a low precision for the flexible human annotation. No comparison between different human flexible annotations were made, and thus the influence of specific human choices are not known. This study applied only to a short list of coarse categories.

Snover et al. (2006) investigated inter-annotator agreement on a post-editing task. They found that even though

the correlation between two human annotators was relatively low, it was still possible to use the annotations from either annotator to calculate a human-targeted error rate.

There are also meta-analytic studies of inter-annotator agreement. Bayerl and Paul (2011) performed a meta-analysis of studies reporting inter-annotator agreement in order to identify factors that influenced agreement. They found for instance that agreement varied depending on domain, the number of categories in the annotation scheme, the training received by the annotators, and calculation method. As no studies on inter-annotator agreement have previously been performed on the domain of MT error analysis, we think that it is important to establish what can be a reasonable level for this task.

3. Annotation experiment

In this section we describe the setup of our error analysis experiment, the error analysis tool, the error taxonomy, and the guidelines used.

3.1. Setup

We have performed error analysis of the output of two different English-Swedish statistical MT systems, trained on Europarl (Koehn, 2005). Both are standard systems, built using the Moses decoder (Koehn et al., 2007), and one of the systems has additional modules for compound processing (Stymne and Holmqvist, 2008). Previous evaluations has shown a small advantage of the system with compound processing, which perform significantly better on a standard Europarl 2000-sentence testset as measured by automatic metrics, as shown in Table 1. Significance was tested using approximate randomization (Riezler and Maxwell, 2005) with 10000 iterations and $p < 0.05$.

The error analysis was performed by the two authors, who are both native Swedish speakers and fluent in English. The annotations were performed using an error annotation tool with which both annotators were well acquainted.

The analysis was done in two phases, with 50 sentences in each phase. In phase 1, the error taxonomy had been designed and discussed by the annotators, but there were no example-based guidelines. Before phase 2 the annotators discussed problematic examples from phase 1, and wrote down guidelines based on that, which were then used in phase 2.

We analysed sentences from the two systems on two sets of sentences. The short set contained sentences with a maximum length of 20 words and average length of 12.1 words. The random set contained random sentences from the testset, with a maximum length of 58 words and average length of 21.9 words.

3.2. The annotation tool

The annotations were performed using the BLAST error annotation tool (Stymne, 2011). The system has three working modes for handling error annotations: for adding new annotations, for editing existing annotations, and for searching among annotations. In annotation mode, the system displays the source sentence, the system translation and a reference translation (if available). Using the system's preprocessing module, similarities between the MT output and the reference are high-lighted with different colouring schemes.

	Bleu	**Meteor**
−*Compound*	21.63	57.86
+*Compound*	22.12	58.43

Table 1: Metric scores

The system is modularized so that error typologies can be added and changed. A typology may have arbitrary depth, for instance sub-classifying error types such as missing or extra words according to the affected part-of-speech and morphological properties. Annotations are stored with the indices of the words they apply to, which enables the production of confusion matrices and a number of statistical computations.

3.3. Error taxonomy

We wanted to use a detailed error taxonomy in our work, to cover as many aspects of system performance as possible. At the same time this would give us data on what distinctions we could make reliably. The major dimensions used are summarized in Table 2. By combining values from all dimensions, we obtain thousands of categories. When analyzing the annotated files, however, we found that only between 50 and 100 of the possible combinations in the error typology had actually been used for each annotated sentence set. We thus report results on different levels of annotation, focusing on subsets of the classes in the full taxonomy. Annotations for seriousness were only done in phase 2.

3.4. Guidelines

When we compared the annotations from Phase 1 (see Table 3), we found a number of cases where there were deviations. All those cases were looked into in order to find explanations for the deviations and a basis for formulating guidelines. Some of the guidelines were quite general such as:

- If there are several possibilities to construct an acceptable word sequence from a garbled system translation, make as few changes (and hence, errors) as possible.
- If a word is affected by several errors, say, both meaning and form, all of them should be registered.

Guidelines affecting specific words or constructions were also needed:

- If a preposition is translated by a wrong preposition, the error should be classified as a Disambiguation error (Word sense) if the meaning is affected; otherwise it should be classified as Wrong function word.
- A missing hyphen should be classified as an orthographic error.

Altogether, we came up with twenty-three guidelines, sorted under nine different categories, ranging from General guidelines to guidelines affecting particular parts-of-speech and construction types.

Label	Description	Relation to previous work
ER	Error-rate based categories: Missing, Extra, Wrong and Word Order	The basis of error rates such as TER (Snover et al., 2006) and the top level in Vilar et al. (2006) and Popović and Burchardt (2011)
Ling	Linguistic categories, such as orthographic, semantic (sense) and syntactic	Similar to Farrús et al. (2010)
GF	Distinction between grammatical and function words	Used to some extent in Vilar et al. (2006)
Form	Subcategories for morphological categories	
POS+	Classifications of which category the error concerns, mainly part-of-speech based, but also contains other categories such as punctuation	Basis of Flanagan (1994), Elliott et al. (2004)
FA	Judgments of whether the error concerns fluency, adequacy, both or neither	Common distinction in MT evaluation in general
Ser	Judgments of the seriousness of an error on a 4-step scale from insignificant (0) to serious (3)	
Reo	Cause of reordering	
Other	Other distinctions	
Index	The position an error has in the sentence. For most errors the position is marked in the system output, but for errors such as Missing words, the position is marked in the source sentence.	

Table 2: Distinctions annotated for in the taxonomy

4. Results

In this section we present the results from our study as concerns inter-annotator agreement, category confusion, and the comparability of error profiles.

4.1. Inter-annotator agreement

It is not clear how to calculate inter-annotator agreement for the error analysis task, since both the number of identified errors and the classifications and positions of the errors can differ. In this analysis we have focused on how the annotated errors co-occur in each sentence, calculated by

$$Agreement = \frac{2 * A^{agree}}{A1^{all} + A2^{all}}$$

where superscript *all* is the total number of annotations by each annotator, and superscript *agree* is the number of annotations on which the annotators agreed. The analysis is made using different combinations from the error taxonomy.

Table 3 summarizes the agreement between the two annotators in both phases. There is a consistently higher agreement in phase 2, when guidelines were used, compared to phase 1. The gap between the phases is larger when more complex annotations were used than for simpler annotation schemes. The classification into adequacy/fluency and for seriousness was difficult, as shown by the big gap between the level of details that included and excluded this. The gap of 6–9 percentage points in the agreement between level 1 and ER, shows that the subclassification for incorrect words into linguistic categories, are not unproblematic for the annotators. Adding the further categories used at level 2, led to a further reduction of agreement. It can be noted that the agreement at level 2 in phase 2 is higher than for level 1 in phase 1.

4.2. Category confusion

We further investigated which types of distinctions between error types that were easy or hard to make, by analysing both the errors that the annotators agreed on, and those which they disagreed on. All the data in this subsection is based on phase 2, where we used guidelines.

	Phase 1	Phase 2
All	–	25%
All-Index	–	28%
All-Ser	27%	40%
All-Ser-Index	32%	45%
All-Ser-Index-FA	39%	58%
ER+Ling+Form+GF+Reo (level 2)	62%	71%
ER+Ling (level 1)	68%	74%
ER	77%	80%
Index	58%	66%
Total number of errors	473	400

Table 3: Inter-annotator agreement in the two annotation phases. Percentage of marked errors that both annotators agreed on, at different levels of annotation (labels refer to those in Table 2)

Confusion matrices for Level 2 errors, excluding distinctions between word order, are shown in Table 4, for the errors for which the annotators marked the same indices. For a large majority of errors the annotators either agree on the classification, 266 cases, or the error is only identified by one annotator, 287 cases.

It is rare that the two annotators have marked the same indices but chosen a different error category. For the errors in Table 4 there are only 44 such cases, which is 7.3% of the total of 597 errors with identical indices. It is also the case, that for the mis-matching errors, the differences are mostly within the top categories, for instance, a Wrong error of some type is mostly confused with a Wrong error of another type, not with Missing, Extra or Order errors. The types of error that are most often confused with other categories are Wrong syntax and sense. Popović and Burchardt (2011) mentioned that the choice between incorrect lexical

	Eco	Egr	Mco	Mgr	O	Wfw	Wfo	Wor	Wse	Wst	Wsx
Eco	2										
Egr	1	5									
Mco	–	–	10								
Mgr	–	–	3	7							
O	–	–	–	–	17						
Wfw	–	–	–	–	–	14					
Wfo	–	–	–	–	–	–	93				
Wor	–	–	–	1	–	1	1	61			
Wse	–	–	–	–	3	–	4	–	25		
Wst	–	–	–	–	1	–	2	–	2	2	
Wsx	–	1	–	–	1	2	4	1	16	–	30
None	6	6	36	22	20	–	42	16	65	31	43

Table 4: Confusion matrix for the categories at Level 2, excluding word order subclassifications. The abbreviations of category names, refer to the categories in Table 7.

	Eco	Egr	Mco	Mgr	O	Wfw	Wfo	Wor	Wse	Wst	Wsx	Tot
Same index	2	5	10	7	17	14	93	61	25	2	30	266
Different index	2	5	22	11	21	14	102	62	40	5	38	322

Table 5: Number of errors for which the annotators agree when the index of the error is taken into account, and when it is not.

choice and missing or extra words is "especially difficult". This is not the case in our evaluation where there is no confusion between Wrong sense, which roughly corresponds to incorrect lexical choice, and Missing or Extra words.

The fact that there are a large number of errors only identified by one annotator can to some extent be explained by the fact that the annotators could have marked different indices. Table 5 shows a comparison of the errors both annotators agreed on when we also count matches with different indices. The number of errors the annotators agreed on increases by over 20% by allowing agreement when the indices do not match. The change is different for different categories, however. For extra words, foreign words and orthographical errors there is hardly any difference, which means it was easy to identify on which words the error occured. For Missing content words and Wrong style errors the indices are mis-matching in more than half of the occurences. In this case our findings are consistent with the discussion in Popović and Burchardt (2011), who also found it hard to decide which words that are part of a reordering, even though this is not the most problematic category in our study. We think it is important to add better guidelines for the placement of indices for the problematic categories.

We also wanted to investigate the difficulty of the different distinctions shown in Table 2 that were made in the error taxonomy. Table 6 shows a summary of how many times the annotators agreed and disagreed for each type of distinction on the errors for which they had marked the same indices. There are some errors for all types of distinctions, but the most problematic distinctions were for adequacy/fluency and seriousness. For the 94 errors for classification between fluency and adequacy, most of them, 76.6% confused either adequacy or fluency with the both classification; it was relatively rare that adequacy and fluency were confused, 9.6%. For seriousness there were no explicit guidelines, and thus the annotators were not especially consistent. In only 6.7% of these errors were the difference more than one step on the 4-step scale, however. For the other categories, the errors were mostly spread out between categories, and there were no clear patterns. The only cases where a specific mismatch stood out was for the linguistic categories, where the distinction between syntax and sense was most common, with 48.5% of the instances.

	Agree	Dis-agree
ER	293	7
Ling	225	33
GF	24	4
Form	78	14
POS+	82	15
FA	214	94
Ser	187	120
Reo	11	6

Table 6: Number of times that the annotators agreed and dis-agreed on a classification for a subcategory (labels refer to those in Table 2)

4.3. Error profiles

We also wanted to see if we could say anything about the error profiles of the two different systems, even though the inter-annotator agreement was relatively low. In these analyses we used details at level 1 and 2, see Table 3. In Table 7 we show the error profiles from phase 2, for comparisons between translation systems, annotators and sentence sets. To investigate whether the error profiles were significantly different we applied chi-square tests to each pair of profiles in both phases; the critical values for these tests are shown in Table 8.

Using a 5%-level of significance, the difference between

	System		Annotator		Sentence length	
	−Compound	+Compound	1	2	short	random
Extra, content	5	6	7	4	5	6
Extra, grammatical	7	15	10	12	6	16
Missing, content	20	40	33	27	17	43
Missing, grammatical	22	20	13	29	10	32
Order, V2	7	6	4	9	2	11
Order, Adverb	8	6	9	5	1	13
Order, other	19	13	18	14	9	23
Wrong, foreign	20	10	14	16	13	17
Wrong, form, agreement	53	48	53	48	40	61
Wrong, form, split compound	15	18	14	19	17	16
Wrong, form, other	64	47	56	55	37	74
Wrong, orthography	73	72	77	68	46	99
Wrong, sense	82	65	66	81	54	93
Wrong, style	21	18	15	24	22	17
Wrong, syntax	64	63	53	74	54	73
Total	480	447	442	485	333	594

Table 7: Error profiles for phase 2, level 2 annotations. Level 1 counts are available by merging counts between each pair of vertical lines.

	Phase 1		Phase 2	
	level 1	level 2	level 1	level 2
System, A1+A2	.27	.28	.12	.43
System, A1	.64	.59	.17	.29
System, A2	**.0005**	**.0005**	**.0005**	**.0005**
Annotator	**.0005**	**.005**	.54	.07
Sentence length	**.03**	**.04**	**.0004**	**.01**

Table 8: Chi-square critical values for the chance of error profiles being different. Significant differences at the 5%-level are marked with bold text.

the two translation systems is always insignificant for annotator 1 and when the two annotators are grouped. For annotator 2, on the other hand, the two systems are significantly different, showing that there is an annotator effect. There are some interesting overall differences in the error profiles, for instance the −Compound system has double the amount of untranslated words as the +Compound system, and many more instances of the category Wrong. It also has more errors in total, and while this difference is not significant, it is still substantial, $p = 0.08$ on a binomial test.

The difference between short and random sentences are always significant, which tells us that the errors on short sentences are not representative for the errors on a standard data set. The difference between the annotators is significant in phase 1, but insignificant in phase 2, again showing the need of guidelines for annotation. In phase 2, the two annotators were more in agreement at level 1, using fewer error categories, than at the more detailed level 2, where it is close to the 5%-level of significance.

5. Conclusion

We have performed a MT error analysis study aimed at investigating the inter-annotator agreement of the task and the statistical significance of the resulting error profiles of systems. We have shown that it is possible to get a reasonable inter-annotator agreement either when using a simple error taxonomy, or when using a more detailed taxonomy and a set of guidelines for the annotators. The use of guidelines in phase 2 improved the error analyses compared to phase 1. Our study thus confirmed the findings of Bayerl and Paul (2011) that inter-annotator agreement is increased for typologies with fewer categories and when more training is received, in our case the use of guidelines.

There were still differences in the significance of error profiles for the two annotators, however. We think that it is important that agreement is further improved by measures such as more joint discussion of examples and more detailed guidelines. With such measures we believe that the rates now obtained, 80% for the top level categories, and 74% with added linguistic sub-classification can be improved substantially.

We have also found that differences in error categorization are often caused by different views on how the error should minimally be corrected. This speaks in favour of developing guidelines in terms of minimal corrections and, perhaps also, explicitly storing the intended target hypothesis, as has been done in connection with learner corpora (Lüdeling, 2011).

We have also shown that sets of short sentences can yield different results than randomly chosen sets. This was a rather disappointing result, but one which we think is significant. We thus must be careful about drawing conclusions on sets of short sentences, which are easier to annotate.

While the current data do not reveal consistent significant qualitative differences between the two systems, the error profiles have some definite tendencies, which can be tested further by taking more data into account, and focusing the annotation on the relevant categories.

This study only investigates these issues for one language pair, and for two relatively similar language pair. In future work, we plan to extend this to other language pairs and translation systems.

6. References

Eleftherios Avramidis and Philipp Koehn. 2008. Enriching morphologically poor languages for statistical machine translation. In *Proceedings of ACL*, pages 763–770, Columbus, Ohio, USA.

Petra Saskia Bayerl and Karsten Ingmar Paul. 2011. What determines inter-coder agreement in manual annotations? A meta-analytic investigation. *Computational Linguistics*, 37(4):727–752.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic.

Debbie Elliott, Anthony Hartley, and Eric Atwell. 2004. A fluency error categorization scheme to guide automated machine translation evaluation. In *Proceedings of AMTA*, pages 64–73, Washington, DC, USA.

Mireia Farrús, Marta R. Costa-jussà, José B. Mariño, and José A. R. Fonollosa. 2010. Linguistic-based evaluation criteria to identify statistical machine translation errors. In *Proceedings of EAMT*, pages 52–57, Saint Raphaël, France.

Mary Flanagan. 1994. Error classification for MT evaluation. In *Proceedings of AMTA*, pages 65–72, Columbia, Maryland, USA.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL, demonstration session*, pages 177–180, Prague, Czech Republic.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit X*, pages 79–86, Phuket, Thailand.

Anke Lüdeling. 2011. Corpora in linguistics: Sampling and annotations. In Karl Grandin, editor, *Going Digital, Evolutionary and Revolutionary Aspects of Digitization*, Nobel Symposium 147, pages 220–243. The Nobel Foundation.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318, Philadelphia, Pennsylvania, USA.

Maja Popović and Aljoscha Burchardt. 2011. From human to automatic error classification for machine translation output. In *Proceedings of EAMT*, pages 265–272, Leuven, Belgium.

Stefan Riezler and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at ACL'05*, pages 57–64, Ann Arbor, Michigan, USA.

Alina Secară. 2005. Translation evaluation – a state of the art survey. In *Proceedings of the eCoLoRe/MeLLANGE Workshop*, pages 39–44, Leeds, UK.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human notation. In *Proceedings of AMTA*, pages 223–231, Cambridge, Massachusetts, USA.

Sara Stymne and Maria Holmqvist. 2008. Processing of Swedish compounds for phrase-based statistical machine translation. In *Proceedings of EAMT*, pages 180–189, Hamburg, Germany.

Sara Stymne. 2011. Blast: A tool for error analysis of machine translation output. In *Proceedings of ACL, demonstration session*, Portland, Oregon, USA.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error analysis of machine translation output. In *Proceedings of LREC*, pages 697–702, Genoa, Italy.