Error Classification and Analysis for Machine Translation Quality Assessment

Maja Popović

Abstract This chapter presents an overview of different approaches and tasks related to classification and analysis of errors in machine translation (MT) output. Manual error classification is a resource- and time-intensive task which suffers from low inter-evaluator agreement, especially if a large number of error classes have to be distinguished. Automatic error analysis can overcome these deficiencies, but state-of-the-art tools are still not able to distinguish detailed error classes, and are prone to confusion between mistranslations, omissions, and additions. Despite these disadvantages, automatic tools can efficiently replace human evaluators both for estimating the distribution of error classes in a given translation output, as well as for comparing different translation outputs. They can also facilitate manual error classification by pre-annotation, since correcting or expanding existing error tags requires less time and effort than assigning error tags from scratch. Classification of post-editing operations is more convenient both for manual and for automatic processing, and also enables more reliable assessment of automatic tools. Apart from assigning error tags to incorrectly translated (groups of) words, error analysis can be performed by examining unmatched sequences of words, part-of-speech (POS) tags or other units, as well as by identifying language-related and linguistically-motivated issues. These linguistic categories can be then used to perform automatic evaluation specifically on these units, or to analyse their frequency and nature. Due to its complexity and variety, error analysis is an active field of research with many possible directions for development and innovation.

Keywords Translation quality assessment · Principles to practice · Automatic evaluation · Translation errors · Machine translation · Post-editing

M. Popović (✉)
Department of English and American Studies, Humboldt University of Berlin, Berlin, Germany
e-mail: maja.popovic@hu-berlin.de

J. Moorkens et al. (eds.), *Translation Quality Assessment*, Machine Translation: Technologies and Applications 1, https://doi.org/10.1007/978-3-319-91241-7_7

1 Introduction

Evaluation of MT is an important but difficult task. How good is a given MT output? Is it good enough for a particular task? These simple questions are not easy to answer because there is no single correct translation of a text. If one sentence is translated by several translators, or even by the same translator at different times, several different translations could be produced. One way to evaluate MT is to present the output to bilingual human evaluators who understand both source and target-languages, in order to assign a quality score for a given task, e.g. from 1 (poor) to 5 (perfect). The criteria normally used are adequacy (i.e. meaning preservation), fluency (i.e. grammaticality), overall quality (based on a combination of both), as well as estimated cognitive post-editing effort. Comparing different MT of the same source text can also be performed by ranking (Callison-Burch et al. 2007), i.e. for each output sentence the evaluator should say if version A or version B is better, without assigning any absolute score. Both approaches can also be performed by monolingual evaluators who understand only the target-language, but a correct reference translation should be available in this case.

The availability of reference translations also enables automatic evaluation, normally using a script or program which produces a score based on the similarity between the reference translation and the MT output. This score is usually produced either as a percentage of matched *n*-grams[1] between the reference and the output or as edit distance between them. Since automatic evaluation is significantly faster and cheaper and also more consistent than human evaluation, a number of automatic evaluation metrics (AEMs) have been investigated and used, e.g. BLEU (Papineni et al. 2002) based on word *n*-gram precision, chrF (Popović 2015) based on character *n*-gram F-score, METEOR (Banerjee and Lavie 2005) based on unigram precision, recall and additional linguistic knowledge, or TER (Snover et al. 2006) based on edit distance (see also Castilho et al. in this volume).

Whereas all of these overall scores and better-or-worse ranking decisions represent very valuable information and help in the continuous improvement of MT systems, MT researchers and developers often find it helpful to have additional information about their systems. What are the most serious problems in a translation system? What are the particular strengths and weaknesses of the system? Does a particular modification improve some aspect of the system, although perhaps it does not improve the overall score? Does a worse-ranked system outperform a higher-ranked one in some aspect? Are some types of errors more difficult to post-edit than others? A relationship between these questions and the overall quality scores is not easy to find.

Therefore, error classification and analysis techniques have emerged, identifying and classifying actual errors in a translated text in order to provide a better foundation for decisions about the task at hand, whether related to system development, purchase, or use. Most often, the goal of error analysis is to obtain an error profile

[1] An *n*-gram is a sequence of N words in a text, so for example where N = 3, this is a trigram: a sequence of three words.

for a translation output, a distribution of errors over the defined error classes. Another application is comparison of different translation outputs, i.e. finding error distributions over different translations for each error class. Furthermore, more specific analyses can be carried out, such as relations between particular error types and user/post-editor preferences, the impact of different error types on different aspects of post-editing effort, and so forth.

Similarly to overall evaluation, error classification is by no means a straightforward task. It can be carried out manually, automatically, or using a combined method (semi-automatically). Different sources of information (in addition to the analysed translation output) can be used, such as source-language texts, reference translations or, recently, post-edited translations. Merely defining a suitable set of error classes (an error typology or taxonomy) is a challenging task in itself: which error types are of interest for the given task and how many details are needed? Once the error typology is defined, for a number of erroneous words there may be several possible error classes, and it is often difficult to determine the position of errors, i.e. to decide which exact words are erroneous.

Apart from classification and annotation of erroneous words, error analysis can be carried out by other means, e.g. by analysing words, POS, or other types of sequences which are not matched when comparing translation output with a reference. Another approach is the definition of linguistically-motivated categories in order to perform error analysis and/or automatic evaluation specifically on them.

For all these reasons which contribute to its complexity, error analysis is an active field of research. This chapter presents a variety of error analysis approaches and error typologies which have been used in the MT community, together with the associated advantages, disadvantages, and challenges. It should be noted that, despite the fact that the described approaches focus on analysis of MT output, the methods can also be used for evaluation of human translations (such as those produced by language learners, non-native speakers, non-experienced translators, and others).

2 Manual Error Classification and Error Typologies

The most obvious method for error analysis is to look into the translation output, mark each erroneous word, and assign a corresponding error tag to it. Apart from the analysed translation output, at least one correct text should be given to the annotator: either the original source text, or a reference translation, or both. The influence of different sources of information and different annotator profiles (bilingual vs. monolingual) has been investigated for assigning overall quality scores (Guzmán et al. 2015). Experiments on Spanish-to-English translation outputs showed that monolinguals are slower but more consistent than bilinguals, and that all annotators become slower and less consistent when exposed to additional information in the form of the source-language text. Therefore, the authors advise monolingual evaluators and the use of reference translations alone. For error classification, to the best of my knowledge, no similar study has been carried out to date.

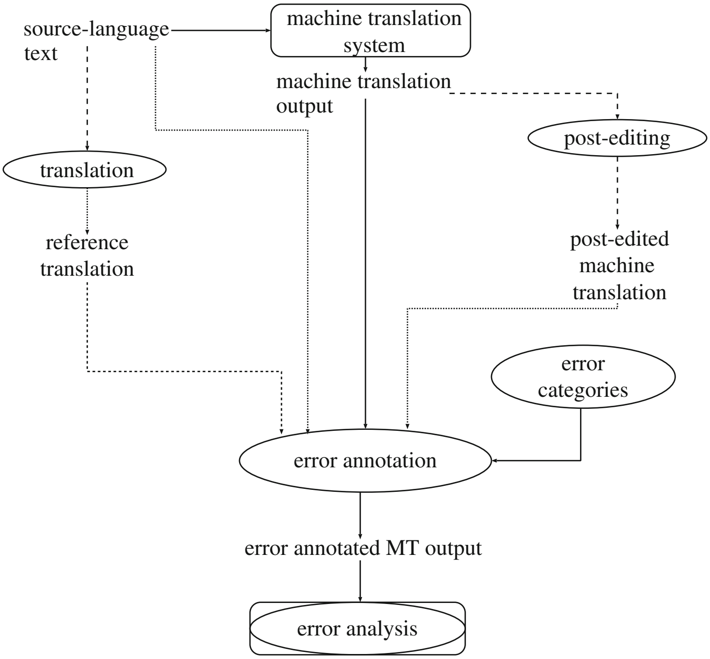

Fig. 1 General procedure of manual error annotation; the rectangle denotes automatic process, and the ellipse denotes manual process

In recent years, post-editing of MT output has become an increasingly common form of human-machine cooperation for translation. Therefore, we have seen more attention given to analysis of post-editing activity through the assignation of an error category to each performed post-edit operation. Usually, the analysis of post-edits is carried out in order to investigate relations between different error types and different aspects of post-editing effort, namely cognitive, temporal, and technical as defined in Krings (2001). For such an analysis, post-edited translation output is necessary as additional information whereas source-language text is optional.

The general process of manual error classification is illustrated in Fig. 1. For any task and approach, a set of error categories (i.e. an error typology or taxonomy) should be clearly defined beforehand. This itself is a demanding task for several reasons: the errors should reflect all advantages and disadvantages of the MT system, which are important for the task at hand as well as for the languages involved; more detailed errors are more informative but more difficult to distinguish; and the error types should cover both linguistic aspects as well as translation aspects. Although there is some work in progress in this direction, there are still no general rules for defining error categories, even on a broad level. The following subsection will present an overview of error typologies for manual classification used in the last decade (i.e. from the beginning until now) for different tasks, including analysis of post-editing process.

Level 1	Level 2	Level 3
Missing words	Content words	
	Filler words	
Word order	Word level	Local range
		Long range
	Phrase level	Local range
		Long range
Incorrect words	Sense	Wrong lexical choice
		Incorrect disambiguation
	Form	
	Extra words	
	Style	
	Idioms	
Unknown words	Unknown stem	
	Unseen forms	
Punctuation		

Table 1 Vilar et al. (2006) error categories

2.1 Overview of Error Typologies and Tasks

Vilar et al. (2006) report the first shift towards the use of explicit error classification and analysis. Error analysis of several Chinese-to-English, Spanish-to-English, and English-to-Spanish statistical MT (SMT) systems is carried out in order to identify the main problems with these systems. The proposed classification scheme presented in Table 1 has a hierarchical structure and is based on the error typology used for refinement of rule-based systems in Llitjós et al. (2005).

Since then, a number of error classification schemes have been used for distinct purposes. The same error scheme is used for error analysis of English-to-Czech MT by Bojar (2011).

Farrús et al. (2010) describe a simple error scheme containing five broad classes (as seen in Table 2) used for comparison of two SMT systems for the Spanish-Catalan language pair in both directions. The systems are also compared in terms of overall human scores, and it is observed that lexical and semantic errors have more influence on human evaluators' perception of quality than other categories. A similar error scheme is used in Comelles et al. (2012) as a basis for development of an AEM based on linguistic features.

Federico et al. (2014) use another similar typology containing a set of basic error classes (see Table 3) for analysing MT from English into Arabic, Chinese, and Russian. For each segment, the annotators marked both erroneous words, as well as assigning an overall quality score using the open-source tool MT-EQuAl[2] (Girardi et al. 2014). These annotations are used to investigate the impact of particular error types and their combinations to the overall quality score using mixed-effect

[2] http://www.mt4cat.org/software/mt-equal

Morphological errors
Lexical errors
Orthographic errors
Syntactic errors
Semantic errors

Table 2 Farrús et al. (2010) error categories

Morphological errors
Lexical choice
Additions
Omissions
Casing and punctuation
Reordering errors
Too many errors

Table 3 MT-EQuAl error categories (Federico et al. 2014)

models (Baayen et al. 2008). The largest correlation is observed for lexical errors and missing words. An additional and very interesting finding is that the human perception of quality does not necessarily depend on the frequency of a given error type; a sentence with a low overall score can easily contain fewer missing words and/or lexical errors than another sentence with a higher score.

A similar, basic typology (without “casing and punctuation” and “too many errors”) was used by Castilho et al. (2017a) and Castilho et al. (2017b) to compare phrase-based SMT and neural MT outputs for a number of language pairs and genres/domains.

Kirchhoff et al. (2012) present another study which examines user preferences regarding different error classes. Different English-to-Spanish translations were annotated with error tags from a detailed typology shown in Table 4 and the overall quality of each translation is estimated by ranking. Then, conjoint analysis[3] is applied in order to find relations. The obtained results also showed that the frequency of a particular error type is not crucial; the least preferred (or most annoying) were word order and word sense errors, whereas the most frequent morphological errors were ranked as third-least preferred.

Stymne and Ahrenberg (2012), in the first work dealing with inter-annotator agreement for error classification, use a distinct but also detailed hierarchical error scheme presented in Table 5. In addition to inter-annotator agreement, the results for two English-to-Swedish MT systems (with and without compound processing) are presented. The error classes were assigned by two annotators, native Swedish speakers, using the BLAST tool for computer-aided manual error analysis (Stymne 2011). The annotation was carried out in two rounds, with and without guidelines.

[3] A “formal framework for preference elicitation”, normally used for consumer studies in which participants rate or rank products based on a combination of attributes (Kirchhoff et al. 2012).

Level 1	Level 2
Missing words	Content words
	Function words
Extra words	Content words
	Function words
Word order	Local range
	Long range
Morphology	Verbal
	Nominal
Word sense error	
Punctuation	
Spelling	
Capitalisation	
Untranslated	Medical term
	Proper name
Pragmatics	
Diacritics	
Other	

Table 4 Kirchhoff et al. (2012) error categories

Level 1	Level 2
Error rates	Missing words
	Extra words
	Wrong word
	Word order
Linguistic	Orthography
	Semantics
	Syntax
GF	Grammatical words
	Function words
Form	Morphological categories
POS+	Part of speech
	Punctuation
FA	Fluency
	Adequacy
	Neither
	Both
Reo (cause of reordering)	
Index (position of an error)	
Other (other categories)	
Ser (seriousness of an error	

Table 5 Stymne and Ahrenberg (2012) error categories

For the detailed error schemes without guidelines, the rate of agreement reached roughly 25%, and guidelines increased this up to 40%. For simple typologies, agreements are in a range of between 65% (without guidelines) and 80% (guided). Aside from this, the authors report that the annotators often disagree regarding the exact positions of erroneous words. Their results also confirmed some findings reported in a study dealing with general inter-annotator agreement (Bayerl and Paul 2011), namely that the number of categories as well as the intensity/absence of training are very important factors.

The Multidimensional Quality Metric (MQM)[4] is used for another study about inter-annotator agreement (Lommel et al. 2014b, see also Lommel in this volume). The metric aims to provide a general mechanism for describing a family of related error categories which includes evaluation of human translations. The main idea is to have a large set of hierarchical error categories which allows selection of any subset appropriate for the task at hand. The metric is already being used for the evaluation and comparison of MT systems, for example by Lommel et al. (2014a) to compare rule-based and phrase-based systems for several language pairs and domains, and by Klubicka et al. (2017) to compare phrase-based and neural systems for English-to-Croatian.

Inter-annotator agreement is explored using a subset of MQM presented in Table 6 on a set of English-Spanish and English-German translation outputs in all directions generated by different MT systems. All outputs were annotated by several[5] professional translators using the open-source tool translate5.[6] The obtained results confirmed the main findings reported in Stymne and Ahrenberg (2012), based on:

(i) the role of number of error categories,
(ii) the importance of annotator training, as well as
(iii) the importance of the exact positions of erroneous words as perceived by different annotators.

In addition, it is shown that "Mistranslation" and "Terminology" are very difficult to distinguish without very intensive training, and that the "Function words" category is generally rather unclear. The "Word Order" category exhibited a high level of general consistency, but there was also a high degree of positional disagreement.

Costa et al. (2015) use yet another hierarchical typology, presented in Table 7, slightly tailored for Romance languages to compare four different English-to-Portuguese translation systems. It is shown that lexical and semantic errors have most impact on sentence-level ranking. Furthermore, highly ranked sentences clearly exhibit a low number of grammatical errors, but the relationship between grammatical errors and poorly-ranked segments remained unclear. Apart from this, high inter-annotator agreement between two annotators is reported, which

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

[5] three (de-en), four (es-en, en-es), or five (en-de)

[6] http://www.translate5.net/

Level 1	Level 2	Level 3	Level 4
Accuracy	Mistranslation		
	Terminology		
	Omission		
	Addition		
	Untranslated		
Fluency	Grammar	Morphology (form)	
		Part-of-speech	
		Agreement	
		Tense/mood/aspect	
		Word order	
		Function words	Missing
			Extra
			Incorrect
	Register/style	Capitalisation	
	Spelling		
	Typography	Punctuation	
	Unintelligible		

Table 6 MQM error categories used for inter-annotator agreement

Level 1	Level 2	Level 3
Orthography	Punctuation	
	Capitalization	
	Spelling	
Lexis	Omission	
	Addition	
	Untranslated	
Grammar	Misselection	Word class
		Verbs
		Agreement
		Contraction
		Misordering
Semantic	Confusion of senses	
	Wrong choice	
	Collocational errors	
	Idioms	
Discourse	Style	
	Variety	
	Should not be translated	

Table 7 Costa et al. (2015) error categories

contradicts the results from former studies. The most probable factor is their removal of words with position disagreement from the calculations, which increased the agreement between the error types.

Level 1	Level 2
Noun phrase	Determiner
	Noun meaning
	Noun number
	Case
	Adjective
Verb phrase	Verb agreement
	Verb meaning
Preposition change	
Co-reference change	

Table 8 Blain et al. (2011) error (edit) categories

Level 1	Level 2	Level 3
Word form change		
Word change		
POS change		
Deleted (insertion)		
Added (omission)		
Order	Phrase level	Distance 1
		Distance ≥2
	Word level	Distance 1
		Distance ≥2

Table 9 Koponen (2012) error (edit) categories

2.1.1 Classification of Post-edit Operations

Blain et al. (2011) use the error scheme presented in Table 8 to analyse two post-edited English-to-French MT outputs from statistical and rule-based systems in the technical domain. After post-editing, the changes, defined as post-editing actions, were classified according to the given typology. Apart from the human classification, automatic classification based on TER (Snover et al. 2006) edit operations and linguistic rules was proposed. Both classification methods revealed that changes were mostly performed on noun meaning, indicating problems with terminology for both MT systems.

Koponen (2012) presents another type of edit operation analysis on English-to-Spanish MT from the news domain. The data set used contains human estimates of post-editing effort which do not necessarily correlate with the actual technical effort (i.e. the number of post-editing operations). In order to explore these differences, segments with high, medium, and low predicted effort were selected and edit operations were annotated according to the error scheme presented in Table 9. The results showed that reordering operations correlate with high predicted effort whereas morphological corrections correlate with low predicted effort. In addition, it is shown that segment length plays a significant role for predictions of post-editing effort regardless of the amount and type of operations that need to be performed, i.e. longer segments tend to be generally perceived as more difficult to post-edit.

Zaretskaya et al. (2016) use a variant of the MQM scheme for a similar analysis on English-to-German translation outputs where post-editing time and cognitive effort are measured for different types of edit operations. It is confirmed that for a number of error types these two aspects do not correlate, e.g. the estimated effort for reordering edits is high but the time is relatively short. Another important finding is that errors involving different types of multi-word expressions are associated with high cognitive and temporal effort.

2.2 Challenges and Possibilities for Facilitation

The previous section has shown that error classification has a large scope of distinct applications and can answer a number of questions that are important for the improvement and development of MT systems, as well as for better understanding of human evaluation criteria and the post-editing process. It has also shown that this useful task is rather time- and resource-intensive, and full of very challenging sub-tasks. Some of the particular challenges are discussed in this section.

Annotator's Profile Annotators can be fluent in both source and target-languages or only in the target-language, in which case a correct reference human translation is needed. To the best of my knowledge, differences between annotator profiles regarding error classification speed, consistency, and performance have not yet been investigated.

Consistency Regardless of the annotators' background, precise guidelines and intensive training are necessary in order to achieve sufficient inter-annotator agreement and to obtain reliable results. The training may have to be carried out in several phases in order to yield an acceptable classification performance. However, even in optimal scenarios, it is not possible to completely avoid certain inconsistencies. One problem is differing perception of the exact positions of erroneous words. This is especially problematic for word-order and phrase-order errors. Another problem is different perception of certain error classes in certain contexts, similar to the problem regarding several correct translations of the same sentence. This type of inconsistency is strongly related to the number and definition of the error categories.

Number of Error Classes More detailed error typologies usually provide better information about the errors, for example separating "morphological error" into "inflectional error", "derivational error", and "compositional error", or using an even deeper hierarchy, such as extending "verb inflection error" into "person", "tense", and "mood". In contrast, more error categories require more cognitive effort for the annotators and also lead to lower consistency and poor inter-annotator agreement. Nevertheless, not only the number, but also the exact definition of error categories is very important. For example, even a simple distinction between adequacy and fluency can be difficult because of certain types of grammatical errors. Usually, all grammatical errors are considered as fluency errors, i.e. they are considered not to have anything to do with the source-language and meaning preservation. However, a number of these errors actually occur due to the properties of the source-language and its differences with the target-language, such as incorrect, missing, or added prepositions, conjunctions, and determiners. Therefore, the exact definition of each error class also plays a significant role in the difficulty of the classification task and reliability of the obtained results.

Definition of Error Classes This mainly depends on the task, but can also depend on the language pair(s). For example, if the texts are drawn from a specific domain, “terminology error” is a very important class, whereas for general domain texts it is not. Similarly, if a Romance language is involved, “verb inflection error” is usually used. Generally, some error classes are more problematic for annotators than others. Whereas an “inflection error” usually does not pose problems, its subcategory “agreement error” often requires high cognitive effort in order to be distinguished from general inflection. Disambiguation between “mistranslation” and its subclass “terminology error” has also been found to be rather difficult without intensive training.

Sometimes an error definition is appropriate for one task and language pair, but becomes insufficiently precise when ported to another language or task. For example, “POS error” is equivalent to “derivational morphology error” for English, but not for German where a large portion of derivations consist of adding different prefixes to verbs. Thus, if a German verb prefix is incorrect, this cannot be tagged as “POS error”.

The following lines of work can help with overcoming some of the described obstacles and facilitate the general process:

(i) unification and generalisation of error typologies,
(ii) annotation of post-edited operations instead of raw translation outputs, as well as
(iii) automation of (a part of) the error classification process.

Unification and Generalisation of Error Typologies Establishing a general error typology which can easily be adapted to different tasks and language pairs could significantly reduce effort and inconsistencies related to the definition of error typology and particular classes. Current work in this direction consists of developing the MQM metric described in Sect. 2.1 (see Lommel, this volume), which offers a very large detailed error set containing several subsets. The idea is to use this set as a starting point and select a desired subset appropriate for the task at hand. A further advantage here would be consolidation with human translation evaluation.

Generalisation can also be achieved in other ways, for example, by using a generalised small set of broad error classes as a starting point and enabling its expansion in distinct directions and depths depending on the task/language pair/domain. For example, it can be observed that certain types of broad error classes are present in one way or another in all typologies described in the previous section: lexical errors, morphological errors, syntactic errors, semantic errors, and

Level 1	Level 2	Level 3
Lexis	Mistranslation	Terminology
	Addition	
	Omission	
	Untranslated	
	Should not be translated	
Morphology	Inflection	Tense, number, person
		Case, number, gender
	Derivation	Part of speech
		Verb aspect
	Composition	
Syntax	Word order	Range
	Phrase order	Range
Semantic	Multi-word expressions	
	Collocations	
	Disambiguation	
Orthography	Capitalisation	
	Punctuation	
	Spelling	
Too many errors		

Table 10 A possible general error typology which starts from broad classes and enables various possibilities for expansion

orthographic errors. A possible general typology on this basis is presented in Table 10, together with a set of suggested expansions. It should be noted that the category "too many errors" should be used carefully: it can be very useful for very low quality segments where errors are really difficult to classify, but on the other hand, backing off to this class should not be overused.

Annotating Post-edit Operations Post-editing and error classification are usually observed and carried out as two separate tasks. Error classification has been carried out on post-editing operations mainly in order to better understand different aspects of the post-editing process, but rarely to analyse properties of an MT system. However, the two tasks are actually highly related; post-editing can be viewed as implicit error annotation, since each edit operation is actually a correction of a translation error. Therefore, merging these two tasks can give better insight into the nature of errors. In addition, it can facilitate the annotation process (whatever is changed should be annotated) and improve inter-annotator agreement by reducing error position inconsistencies.

Automatic Error Classification An obvious method for reducing efforts of manual error classification is automation of the process. The advantages and disadvantages are similar to those of automatic evaluation metrics, i.e. faster, cheaper and more consistent, but also less precise, prone to assignment of incorrect error tags, and strongly dependent on a given reference translation. A detailed overview of automatic error classification is provided in the next section.

3 Automatic Error Classification

As mentioned in the previous section, automatic methods for error classification emerged due to resource – and time – intensity, as well as the inconsistency of the manual process. The motivation is the same as for AEMs, namely use a program to compare the translation output with a reference translation. The goal, however, is not to produce a single overall score, but to estimate the amount of different error types.

One of the first steps in this direction (Popović et al. 2006) proposes automatic estimation of reordering and inflectional errors based on Word Error Rate (WER) and Position-independent Word Error Rate (PER) differences. WER, i.e. word level Levenshtein distance (Levenshtein 1966), requires exactly the same order of words in hypothesis and reference segments. PER, on the other hand, neglects word order completely and measures only the difference in the count of words occurring in hypothesis and reference segments. For both metrics, the resulting number of errors is divided by the number of words in the reference. The main idea is that the reordering errors are reflected in the difference between WER and PER, and the inflectional errors are correlated with the difference between PER of original words and PER of lemmas. More detailed analysis of Spanish verb inflections based on the same approach is described in Popović and Ney (2006).

Estimating the amount of inflectional errors and omissions by identification of actual erroneous words contributing to WER and PER is described in Popović and Ney (2007). Further work in this direction resulted in a complete automatic classification scheme (Popović and Ney 2011), which covers a large portion of broad error classes used in human error analysis. These are:

- inflectional errors,
- reordering errors,
- missing words (omissions),
- extra words (additions), and
- lexical errors (mistranslations).

The word-level alignment between the translation output and the reference translation is based on WER, and precision and recall are used as additional information for classification of erroneous words. The transition from PER to precision and recall emerged from the inability of the standard efficient algorithms for PER to give precise information about contributing words. Therefore alternative PER-based metrics were introduced – HPER, RPER, and FPER – which basically correspond to the precision, recall and F-score. The open-source tool Hjerson[7] (Popović 2011) is based on this scheme. The original version of the tool required lemmas in addition to the original word forms in order to distinguish inflectional errors. The extended version (Hjerson+ as described in Popović et al. 2015) enables back-off to the first four characters of the word if the lemmas are not available.

[7] https://github.com/cidermole/hjerson

Another automatic error classification tool with a similar but slightly larger set of error classes is Addicter[8] (Fishel et al. 2011; Zeman et al. 2011). In addition to the previously described five Hjerson error classes, the tool allows detection of untranslated words as well as a variable span of reordering errors (short and long range). The word-level alignment is based on a first-order Markov dependency model, similar to bilingual Hidden Markov Model (HMM)-based word alignment used for MT (Vogel et al. 1996). It stimulates adjacent words to be aligned similarly, which results in a preference towards aligning longer phrases. The tool also accepts external alignments (from GIZA++ (Och and Ney 2003), METEOR, etc.). Lemmas are also required for distinguishing inflectional errors from lexical errors.

Similarly to AEMs, automatic error classification also suffers from relying on just one of many viable reference translations. Therefore both tools accept multiple references: for each segment, Hjerson chooses the reference with minimal WER, and Addicter chooses the reference with the minimal total number of errors. The general procedure for automatic error classification is presented in Fig. 2.

Both tools were tested by comparing the results with those obtained by manual error classification and exhibited sufficiently high correlations to be able to replace human annotators for a number of tasks. The details of automatic error-classification assessment process will be described in the next section.

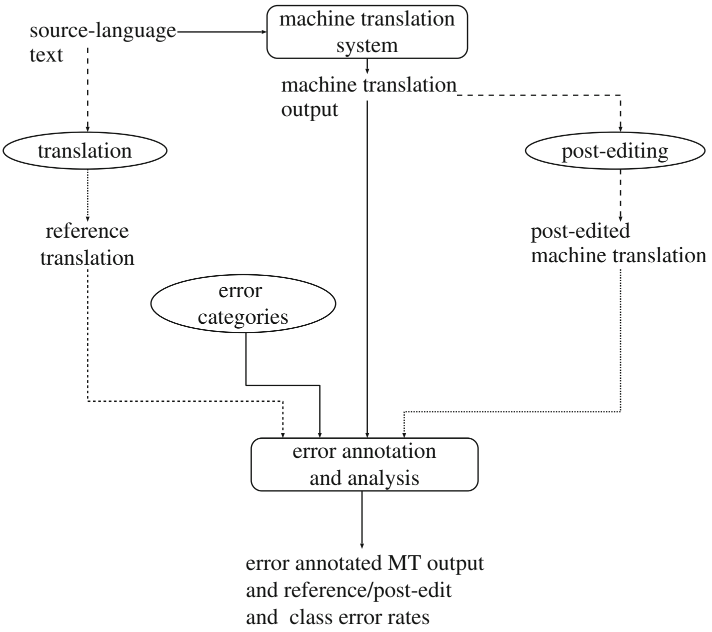
source-language text
machine translation system
machine translation output
translation
post-editing
reference translation
post-edited machine translation
error categories
error annotation and analysis
error annotated MT output and reference/post-edit and class error rates

Fig. 2 Procedure of automatic error classification; the rectangle denotes automatic process, and the ellipse denotes manual process

[8] https://wiki.ufal.ms.mff.cuni.cz/user:zeman:addicter

3.1 Evaluation of Automatic Error Classification

In principle, evaluation of automatic error classification consists of comparing results with the results of manual classification for the following three aspects:

1. Distribution of different error classes within a translation output,
2. Distribution of an error class across different translation outputs,
3. Detecting actual erroneous words and assigning a correct error tag.

For the first two aspects, Spearman and Pearson correlation coefficients between manual and automatic scores are calculated. For the third aspect, for each error class, precision and recall are calculated together with the percentage of confusions with each of the remaining classes.

Assessment of Hjerson and Addicter Both automatic evaluation tools, especially Hjerson, exhibit high correlations for the first as well as for the second aspect, and they are already being used for obtaining error profiles or comparisons of MT systems as well as for some other analyses. As for the third aspect, high recall has been reached for all error classes. Nevertheless, precision scores are rather low, mainly because the number of automatically-detected errors is generally much higher due to the usage of one reference translation; many detected errors are not real errors but just correct variations.

Another disadvantage is the high degree of confusion between lexical errors, omissions, and additions. This distinction, however, is often problematic even for human annotators. Another similarity to manual classification is frequent disagreement regarding position of reordering errors, which decreases both inter-annotator agreement for the manual process as well as precision and recall for automatic tools.

Drawbacks of the Assessment Method Comparison with the results of manual error classification is the most natural way to assess automatic tools, but it should be taken into account that the exact process of manual annotation can influence the results. First of all, the information which was available to the annotators plays an important role; if the reference translation is available, the results of human and automatic classification will be closer than if only the source text is used. Furthermore, if only the reference translation is used, without the source text, the results will be even closer. The annotation guidelines are also important; if the annotators were told to pay specific attention to the reference, the results will be closer than if the reference was used only for orientation.

Another important factor is the fact that the vast majority of manual error-classification tasks have not been carried out for the sake of evaluation of an automatic tool; for a small number of tasks when that was the case, the results are closer. Furthermore, since there is no general error typology for manual error classification, exact mapping to a narrower automatic error typology also differs from task to task. Due to all these factors, the automatic tools available so far had to be evaluated on rather heterogeneous data. These annotated texts were eventually collected, partially homogenised and published as the Terra corpus[9] (Fishel et al. 2012), and despite the described disadvantages, represent a valuable corpus for further development of automatic error-classification tools.

Evaluating on Post-edited Data Recently, the Hjerson tool has been applied for automatic analysis of post-editing operations, such as exploring relations between different error (or edit) classes and different aspects of post-editing effort (cognitive, temporal, and technical; Popović et al. 2014). The results confirmed the main findings of Koponen (2012), and showed that sentence length, in addition to cognitive effort, strongly influences temporal effort. It is also shown that technical effort for all edit classes strongly correlates with estimated cognitive effort regardless of temporal effort. In the experiments, it is observed that automatic error classification produces more reliable results when post-edited MT output is used as a reference translation. Therefore, systematic experiments have been carried out in this direction and the details are presented in the next section.

3.2 Semi-automatic Classification of Post-edit Operations

As mentioned in Sect. 2.2, post-editing and error classification are closely related tasks since post-editing can be viewed as implicit error annotation. Therefore, classification of post-editing operations can not only facilitate manual error classification, but also enable more reliable automatic error classification. In addition, it can also provide more reliable assessment of automatic tools and give a better insight into possibilities for their improvement.

These premises are thoroughly investigated by Popović and Arcan (2016). In their study, a set of around 2800 segments containing different language pairs were post-edited, annotated, and analysed, thus creating a publicly available resource.[10] The texts are first post-edited, then the error annotation is performed in two stages in order to facilitate the manual part: the first stage consists of automatic pre-annotation by Hjerson, and the second stage consists of correcting or expanding Hjerson error classes by human annotators. In addition to the five Hjerson classes, three additional error classes were introduced based on findings in the data:

(i) contraction errors, including any merging of words, mainly compounds,
(ii) derivational morphology errors, and
(iii) untranslated words.

[9] http://terra.cl.uzh.ch/terra-corpus-collection.html

[10] http://nlp.insight-centre.org/research/resources/pe2rr/

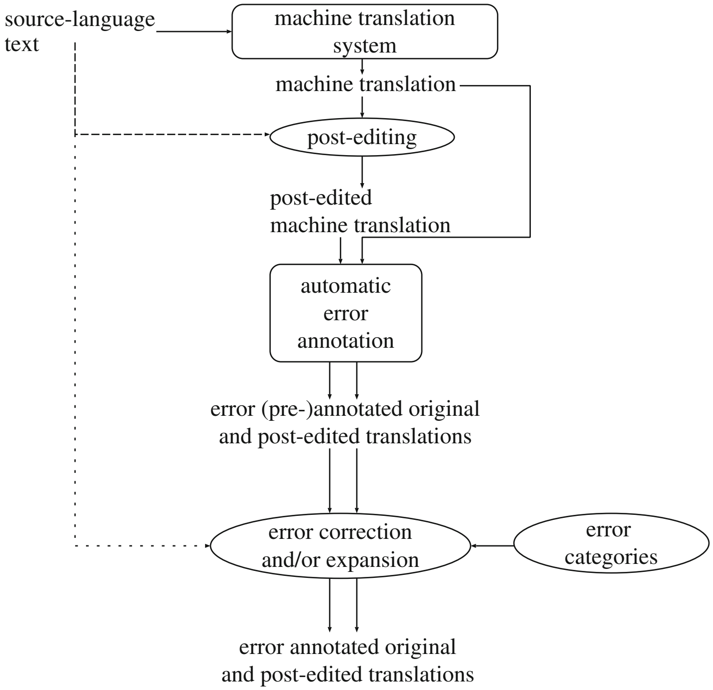
source-language text
machine translation system
machine translation
post-editing
post-edited machine translation
automatic error annotation
error (pre-)annotated original and post-edited translations
error correction and/or expansion
error categories
error annotated original and post-edited translations

Fig. 3 Procedure of manual error annotation of post-edit operations using automatic pre-annotation; the rectangle denotes automatic process, the ellipse denotes manual process

In addition, multiple error tags are assigned when necessary, mainly for reordering errors which are also incorrectly translated in some way. The general procedure for such error annotation is presented in Fig. 3. Rectangular processes were carried out automatically and elliptical processes manually.

When the Hjerson tool was tested on this corpus and results compared to the previous assessment, the correlations for error distributions remained high, and precision improved significantly. Recall either improved or remained unchanged. This confirmed the hypothesis that annotated post-editing operations are more suitable for assessment and development of an automatic error-classification tool. As for Hjerson itself, despite a large improvement in precision, a significant (albeit smaller) amount of confusion between lexical errors, omissions, and additions still remains. Therefore, addressing this problem should be one of the first steps for its improvement.

Taking into account the described findings regarding confusions between lexical errors, omissions and additions, Bentivogli et al. (2016) and Toral and Sánchez-Cartagena (2017) used automatic classification for three broad error types for comparing phrase-based and neural MT outputs: morphological errors, reordering errors and lexical errors, which also comprise additions and omissions.

4 Other Methods for Error Analysis

Apart from explicit and implicit error classification through assignment of tags to erroneous/edited words, other approaches also enable better understanding of the advantages and problems of MT systems, such as identification and analysis of unmatched patterns, as well as checking and evaluating specific linguistic features.

4.1 Analysis of (Un)matched Sequences

The basis for such analysis is automatic comparison of the translation output with a reference translation and detecting either "recall" mismatches, i.e. sequences in the reference segment that are not present in the output segment, or "precision" mismatches, i.e. sequences in the translation output segment that are not present in the reference. Further analysis of these patterns can be carried out either automatically or manually.

Automatic analysis of POS sequences in translation output is proposed by Lopez and Resnik (2005) in order to see how well a translation system is capable of capturing systematic reordering patterns. Recall is calculated for every POS sequence in a translation output, and the patterns with a low recall score are considered as problematic.

The publicly available evaluation tool rgbF (Popović 2012) which calculates n-gram precision, recall, and F-score also enables detecting unmatched n-grams for arbitrary units: words, POS tags, lemmas, morphemes, etc. The tool provides a list of unmatched n-grams, both in the translation output (precision) as well as in the reference translation (recall). Further analysis is left to the user, depending on the task and goals.

Another open-source tool MT-ComparEval[11] (Klejch et al. 2015), for comparing and evaluating different MT systems by several measures, also offers n-gram matching. The tool identifies both unmatched as well as confirmed n-grams (those appearing both in the translation output as well as in the reference segment). When comparing two translation outputs, the tool provides information about improving n-grams (i.e. confirmed n-grams occurring in only one of the outputs), as well as worsening n-grams (i.e. unmatched n-grams occurring in only one of the outputs).

[11] https://github.com/choko/MT-ComparEval

4.2 Evaluating Specific Linguistic Phenomena – Linguistic Check-Points

Another approach for error analysis is to define specific linguistic units, such as a noun phrase, an ambiguous word, or a verb-object collocation, and perform evaluation specifically on them. The general method is to divide each segment into a collection of sub-units which can be classified into linguistic categories and evaluated separately.

First, a linguistic check-point database has to be created from a parallel bilingual text. Both source- and target-language sentences have to be parsed and then linguistic units for each of the defined linguistic categories have to be identified in the parsed sentences. Linguistic units on the target side are directly used as reference translations whereas those on the source side have to be mapped into the target-language. This mapping can be carried out by automatic or manual word alignment and/or other knowledge resources, such as dictionaries or manually-defined rules. Extracted linguistic units and their reference translations represent the linguistic check-point database.

Once a database is available, the evaluation is performed in the following way:

- source sentences containing the desired linguistic categories are selected and translated by an MT system;
- for each check-point, the percentage of matched reference *n*-grams (i.e. recall) is calculated;
- the total score for the given linguistic category is obtained by summing up the scores of all detected check-points.

Zhou et al. (2008) first proposed this approach to analyse problems in English-Chinese MT output using the Woodpecker tool (Wang et al. 2014). Toral et al. (2012) developed the DELiC4MT tool,[12] which builds on the concept introduced by Woodpecker and overcomes two of its limitations:

(i) DELiC4MT is language-independent, while Woodpecker is designed for English-Chinese, and adaptation to other language pairs is not straightforward,
(ii) DELiC4MT's licence allows anyone to work on it and release modifications, while Woodpecker's licence (MSR-LA) is quite restrictive in this regard.

The general procedure for this type of evaluation is shown in Fig. 4. Similar to the standard error-classification task where the error typology has to be defined, one of the important steps for the linguistic check-point approach is the definition of linguistic categories. Such a linguistic typology is, in principle, an inventory of linguistic phenomena of the source-language that can present problems due to, for example, inherent ambiguity, or for translation into a specific target-language, for instance because of syntactic divergence between the two languages involved in

[12] http://www.computing.dcu.ie/~atoral/delic4mt/

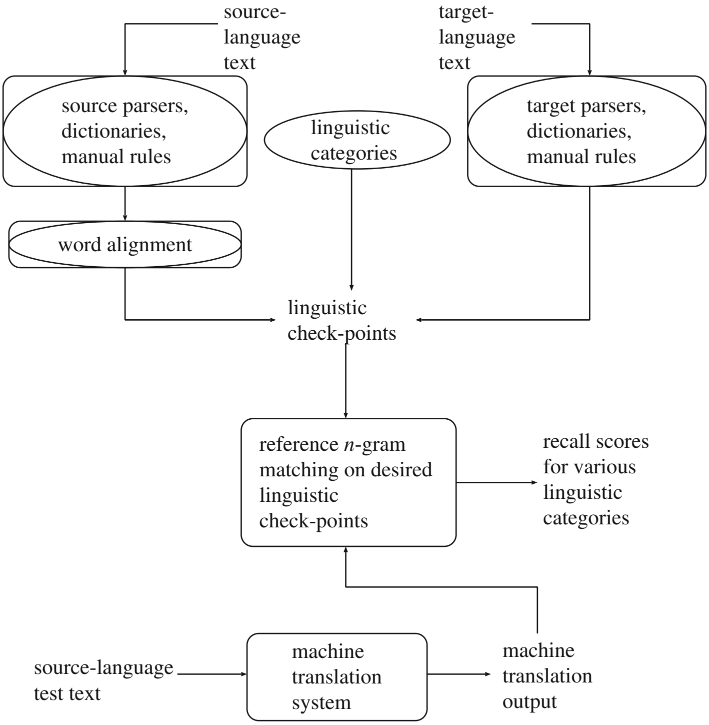
source-language text
target-language text
source parsers, dictionaries, manual rules
linguistic categories
target parsers, dictionaries, manual rules
word alignment
linguistic check-points
reference n-gram matching on desired linguistic check-points
recall scores for various linguistic categories
source-language test text
machine translation system
machine translation output

Fig. 4 Procedure of evaluation on linguistic check-points; the rectangle denotes automatic process, and the ellipse denotes manual process

the translation process. The level of detail and the specific linguistic phenomena included in the typology can vary, depending on what the developers and/or the end-users want to investigate as part of the diagnostic evaluation and on the number of aspects that they are interested in.

The Woodpecker linguistic typology is presented in Table 11. It is based on rich linguistic knowledge from various resources and includes important language phenomena on different linguistic levels in both English and Chinese. Different categories are defined by means of different information sources: some by POS tags, others by dependency tags, others by the use of a dictionary, and some of them, especially those on the sentence level, by manual rules. For DELiCM4T, on the other hand, there is no predefined linguistic typology – the tool enables user-defined language-independent specifications, which are then extracted from texts automatically by Kybots (Knowledge-Yielding Robots) which use a collection of profiles that represent patterns of information of interest (Vossen et al. 2010).

Apart from linguistic check-point evaluation, similar linguistic typologies have been explored for other types of evaluation, which are described in detail in the following section.

	Word level	Phrase level	Sentence level
English	Noun	Noun phrase	Time clause
	Verb	Verb phrase	Reason clause
	Adjective	Adjectival phrase	Conditional clause
	Adverb	Adverb phrase	Result clause
	Preposition	Prepositional phrase	Purpose clause
	Pronoun		
	Modal verb		
	Plural		
	Ambiguous word		
	Possessive pronoun		
	Comparative and superlative		
Chinese	Noun	Subject-predicate phrase	Ba sentence
	Verb	Predicate-object phrase	Bei sentence (passive)
	Adjective	Preposition-object phrase	Shi sentence
	Adverb	Measure phrase	You sentence
	Pronoun	Location phrase	Compound sentence
	Preposition		
	Quantifier		
	Ambiguous word		
	Idiom		
	New word		
	Overlapping word		
	Collocation		

Table 11 Zhou et al. (2008) linguistic categories

4.3 *Identifying and Analysing Language-Related Issues*

Identifying patterns that are causing translation problems due to the characteristics of the involved languages and differences between them can be used not only for linguistic check-point evaluation, but also for analysis of MT systems that goes beyond the standard error classification. For example, phrase-based SMT systems tend to have problems with long-range dependencies involving German verbs. Actual errors that emerged in affected sentences were not only the reordering errors of English verbs, but also missing verbs, as well as mistranslations of other parts of the sentence. The standard error categories for these segments would be “(verb) reordering error”, “missing verb”, and “mistranslation”, and the language-related issue would be “the German verb structure”.

Popović and Arcan (2015) present an identification of such patterns for SMT between Slovenian and Serbian on one side and English or German on the other. The analysis is carried out semi-automatically, namely by manual inspection of texts automatically annotated by Hjerson. Definition of issues is based both on general linguistic knowledge, as well as on phenomena related to the (machine)

Languages	Issues
General	Phrase structure and boundaries
	Literal translations
	Structures involving auxiliary verbs
	Structures involving modal verbs
	Prepositions
	Negation
English	Noun (+adjective) collocations
German, south Slavic	Noun-verb disambiguation
	Inflections (case, gender, number, person, tense)
German	Compositional morphology
South Slavic	Articles

Table 12 Some of the most prominent language related issues found in the PE2rr corpus

translation process. Some of the identified issues are common for all translation directions, whereas some of them depend on the language pair and/or on the translation direction. The PE2rr corpus (Popović and Arcan 2016) described in Sect. 3.2 is partly annotated with these types of issues, and the most frequent ones across different language pairs are shown in Table 12. The same approach is used for analysing issues related to translation between closely-related South Slavic languages (Popović et al. 2016), and for comparison of problematic patterns for phrase-based and neural German-English MT systems (Popović 2017).

Comelles et al. (2016) present a similar study dealing with identification and classification of relevant linguistic features based on general linguistic knowledge as well as on phenomena occurring in a given corpus. The basic typology from Farrús et al. (2010) was extended, as shown in Table 13, and used for development of a linguistically-motivated AEM called VERTa (Comelles et al. 2012), which enables the use of different combinations of the described linguistic features.

Recently, another approach for analysis of language-related phenomena emerged, namely the creation of test sets targeted to specific phenomena. The main advantage over the previously-described approach, which uses "real" data, is the controlled distribution and frequency of desired phenomena.

Guillou and Hardmeier (2016) developed the test suite PROTEST, specifically designed for evaluation of pronoun translation from English to French. The annotation and selection of English source segments is carried out manually, and the evaluation is done manually only for those pronouns which are not found in the reference French translation; an automatic evaluation script is available to discard pronouns that are present in the reference.

Burchardt et al. (2017) developed a corpus containing a larger set of linguistic phenomena for the German-English language pair which was used for comparing three approaches for building MT systems: rule-based, phrase-based, and neural network-based. The selection of language-related phenomena is based on linguistic knowledge, the corpus is created manually, and the translation quality is reported as percentage of correctly-translated instances. A similar strategy is proposed in

Orthography	Capitalisation
	Punctuation
	Date, time, money
Lexical error	Multi-word expressions
	Acronyms and abbreviations
	Untranslated source words
	Omissions
	Proper nouns
Morphology	Inflectional
	Derivational
	Compounding
	Morpho-syntax
Syntax	Syntactic structure
	Word order
	Prepositions
	Relative clauses
	Ungrammatical chunks
Semantics	Lexical semantic relations (synonymy, homonymy, etc.)
	Sentence semantics

Table 13 Comelles et al. (2016) linguistic categories

Isabelle et al. (2017): An English-to-French test corpus of about 100 sentences is created, which contains short examples of several morpho-syntactic phenomena, motivated both by linguistic knowledge as well as experience with issues for phrase-based MT.

Burlot and Yvon (2017) evaluate morphological variations in the target-languages for translation from English to Czech and Latvian. Each segment contains a structure which is expressed syntactically in English but morphologically in the target-language. This work reports the first steps towards automation of the process, using language-model probabilities for the extraction of desired segments (Fig. 5).

4.4 Challenges

After a long hiatus, identification of language-related issues for MT has re-emerged relatively recently and recent works in progress are in the preliminary stages. Therefore, the following important aspects have to be taken into account for further work and development of this evaluation approach:

Definition The decision as to which particular phenomena to concentrate on is far from trivial. The issues have to be linguistically-motivated so that they can reflect

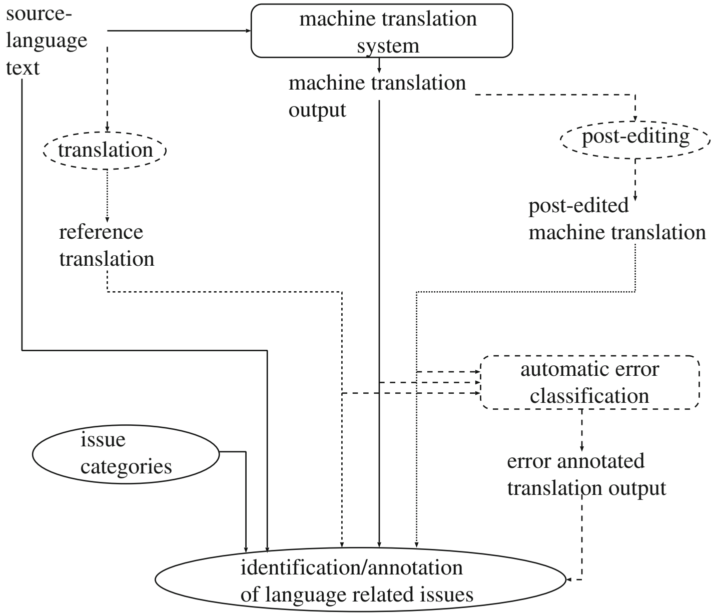
source-language text
machine translation system
machine translation output
translation
post-editing
reference translation
post-edited machine translation
automatic error classification
issue categories
error annotated translation output
identification/annotation of language related issues

Fig. 5 Identification of language-related issues; the rectangle denotes automatic process, and the ellipse denotes manual process

the (in)ability of an MT system to translate specific linguistic phenomena. However, they should not only contain traditional linguistic categories but also categories which are related to the translation process.

Generalisation The issues should be clearly defined and widely accepted so that the results can be easily understood and shared. Similarly to error-typology generalisation, the optimal way would be to establish a broad class of issues which can easily be expanded in different directions appropriate for the languages involved and task at hand.

Relation to Error Categories Although some of the issues defined so far directly correspond to some typical error categories, such as "inflectional error", for a number of issues such a relationship is hard to find. Finding correspondences between the two types of categories can be useful for both tasks.

Automation Analogously to MT evaluation and error classification, automation of issue identification would be beneficial in order to speed up the process and increase consistency. Some of the issues can already be detected automatically but there are a number of directions for future work.

5 Summary

This chapter presents an overview of different approaches and tasks related to the classification and analysis of errors in MT output.

Manual error classification can provide more detail as human annotators can distinguish a larger number of error classes than state-of-the-art automatic tools, but it is a very difficult task for several reasons. The main disadvantages are high costs in terms of time and money, as well as low consistency, especially if the error categories are numerous and complex. In addition, defining an appropriate error typology represents a challenging task itself. Ongoing work (see Lommel, this volume) aims at generalisation by offering a large typology from which an appropriate sub-set could be selected for the task at hand. Generalisation could also start from a set of broad classes and enable different ways and depths of expansion according to the language (pair) and the goal of the evaluation.

Automatic error analysis is faster, cheaper, and more consistent, yet state-of-the-art tools are still not able to provide many details. In addition, existing tools are prone to confusion between certain error classes, although some of these distinctions are not easy even for human evaluators. Despite these drawbacks, automatic classification tools can replace human evaluators both for obtaining an error profile (distribution of error classes) for a given translation output, as well as for comparing different translation outputs. Apart from this, they can facilitate manual error classification by introducing a pre-annotation step; correcting or expanding existing error tags requires less effort and time than assigning error tags to an unannotated text from scratch.

Classification of post-editing operations both by human evaluators as well as by automatic tools is normally used for analysing the post-editing process, and rarely for analysis of translation errors. However, the edit-classification results are more reliable than error classification of raw translation output since the two tasks are actually closely related; post-editing is actually error correction, and therefore can be viewed as implicit error annotation. In addition, annotated post-editing operations are more appropriate for the assessment of automatic classification tools. Of course, post-editing is a resource-intensive task that has to be performed by qualified translators, but taking into account that some kind of human processing is always needed, post-editing certainly represents a good option.

Apart from the typical error classification carried out by assigning error tags to incorrectly translated (groups of) words, other approaches have been used as well. One method is analysis of unmatched sequences of words, POS tags, or other units, such as the sequences which do not appear both in the translation output and in the reference translation. Another approach aims to identify language-related and linguistically-motivated issues in order to automatically evaluate them specifically. Such issues have also been used for analysing their frequency and nature in order to better understand the language-related phenomena that are difficult for an MT system to handle.

Due to its complexity and variety, error analysis is an active field of research with many possible directions for development and innovation. Regarding details about any particular approach or task, all relevant references are given for further reading.

References

Baayen HR, Davidson DJ, Bates DM (2008) Mixed-effects modeling with crossed random effects for subjects and items. J Mem Lang 59(4):390–412

Banerjee S, Lavie A (2005) METEOR: an automatic metric for MT evaluation with improved correlation with human judgements. In: Proceedings of the ACL 05 Workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, Ann Arbor, pp 65–72

Bayerl PS, Paul KI (2011) what determines inter-coder agreement in manual annotations? A meta-analytic investigation. Comput Linguist 37(4):699–725

Bentivogli L, Bisazza A, Cettolo M, Federico Ml (2016) Neural versus phrase-based machine translation quality: a case study. In: Proceedings of the 2016 conference on Empirical Methods in Natural Language Processing (EMNLP2016), Austin, pp 257–267

Blain F, Senellart J, Schwenk H, Plitt M, Roturier J (2011) Qualitative analysis of post-editing for high quality machine translation. In: Machine Translation Summit XIII, Xiamen

Bojar O (2011) Analyzing error types in English-Czech machine translation. Prague Bull Math Linguist 95:63–76

Burchardt A, Macketanz V, Dehdari J, Heigold G, Peter JT, Williams P (2017) A linguistic evaluation of rule-based, phrase-based, and neural MT engines. Prague Bull Math Linguist 108(1):159–170

Burlot F, Yvon F (2017) Evaluating the morphological competence of machine translation systems. In: Proceedings of the 2nd conference on Statistical Machine Translation (WMT 2017), Copenhagen, pp 43–55

Callison-Burch C, Fordyce C, Koehn P, Monz C, Schroeder J (2007) (Meta-)evaluation of machine translation. In: Proceedings of the 2nd workshop on Statistical Machine Translation (WMT 2007), Prague, pp 136–158

Castilho S, Moorkens J, Gaspari F, Calixto I, Tinsley J, Way A (2017a) Is neural machine translation the new state of the art? Prague Bull Math Linguist 108(1):109–120

Castilho S, Moorkens J, Gaspari F, Sennrich R, Sosoni V, Georgakopoulou P, Lohar P, Way A, Barone AVM, Gialama M (2017b) A comparative quality evaluation of PBSMT and NMT using professional translators. In: Proceedings of MT Summit XVI, Nagoya, pp 116–131

Comelles E, Atserias J, Arranz V, Castellón I (2012) VERTa: linguistic features in MT evaluation. In: Proceedings of the 8th international conference on Language Resources and Evaluation (LREC 2012), Istanbul

Comelles E, Arranz V, Castellón I (2016) Guiding automatic MT evaluation by means of linguistic features. Digital Scholarship in the Humanities

Costa A, Ling W, Luís T, Correia R, Coheur L (2015) A linguistically motivated taxonomy for machine translation error analysis. Mach Transl 29(2):127–161

Farrús M, Costa-Jussà MR, Mariño JB, Fonollosa JAR (2010) Linguistic-based evaluation criteria to identify statistical machine translation errors. In: Proceedings of the 14th annual conference of the European Association for Machine Translation (EAMT 2010), Saint-Raphael, pp 167–173

Federico M, Negri M, Bentivogli L, Turchi M (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 2014), Doha, pp 1643–1653

Fishel M, Bojar O, Zeman D, Berka J (2011) Automatic translation error analysis, Pilsen, pp 72–79

Fishel M, Bojar O, Popović M (2012) Terra: a collection of translation error-annotated corpora. In: Proceedings of the 8th international conference on Language Resources and Evaluation (LREC-12), Istanbul, pp 7–14

Girardi C, Bentivogli L, Farajian MA, Federico M (2014) MT-EQuAl: a toolkit for human assessment of machine translation output. In: 25th international conference on Computational Linguistics (CoLing), System Demonstrations, Dublin, pp 120–123

Guillou L, Hardmeier C (2016) PROTEST: a test suite for evaluating pronouns in machine translation. In: Proceedings of the tenth international conference on Language Resources and Evaluation (LREC 2016), Portoroz

Guzmán F, Abdelali A, Temnikova I, Sajjad H, Vogel S (2015) How do humans evaluate machine translation. In: Proceedings of the 10th workshop on Statistical Machine Translation (WMT 2015), Lisbon, pp 457–466

Isabelle P, Cherry C, Foster G (2017) A challenge set approach to evaluating machine translation. In: Proceedings of the 2017 conference on Empirical Methods in Natural Language Processing (EMNLP 2017), Copenhagen, pp 2476–2486

Kirchhoff K, Capurro D, Turner A (2012) Evaluating user preferences in machine translation using conjoint analysis. In: Proceedings of the 6th conference of European Association for Machine Translation (EAMT-12), Trento, pp 119–126

Klejch O, Avramidis E, Burchardt A, Popel M (2015) MT-ComparEval: graphical evaluation interface for machine translation development. Prague Bull Math Linguist 104:63–74

Klubicka F, Toral A, Sánchez-Cartagena VM (2017) Fine-grained human evaluation of neural versus phrase-based machine translation. Prague Bull Math Linguist 108(1):121–132

Koponen M (2012) Comparing human perceptions of post-editing effort with post-editing operations. In: Proceedings of the seventh workshop on Statistical Machine Translation, Montreal, pp 181–190

Krings HP (2001) Repairing texts: empirical investigations of machine translation post-editing processes. Kent State University Press, Kent

Levenshtein VI (1966) Binary codes capable of correcting deletions, insertions and reversals. Sov Phys Dokl 10(8):707–710

Llitjós AF, Carbonell JG, Lavie A (2005) A framework for interactive and automatic refinement of transfer-based machine translation. In: Proceedings of the 10th conference of European Association for Machine Translation (EAMT2005), Budapest, pp 87–96

Lommel A, Burchardt A, Popović M, Harris K, Avramidis E, Uszkoreit H (2014a) Using a new analytic measure for the annotation and analysis of MT errors on real data. In: Proceedings of the 17th annual conference of the European Association for Machine Translation (EAMT 2014), pp 165–172

Lommel A, Popović M, Burchardt A (2014b) Assessing inter-annotator agreement for translation error annotation. In: Proceedings of MTE workshop on automatic and manual metrics for operational translation evaluation, LREC 2014, Reykjavík

Lopez A, Resnik P (2005) Pattern visualization for machine translation output. In: Proceedings of HLT/EMNLP on interactive demonstrations, Vancouver, pp 12–13

Och FJ, Ney H (2003) A systematic comparison of various statistical alignment models. Comput Linguist 29(1):19–51

Papineni K, Roukos S, Ward T, Zhu WJ (2002) BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL 2002), Philadelphia, pp 311–318

Popović M (2011) Hjerson: an open source tool for automatic error classification of machine translation output. Prague Bull Math Linguist 96:59–68

Popović M (2012) RgbF: an open source tool for *n*-gram based automatic evaluation of machine translation output. Prague Bull Math Linguist 98:99–108

Popović M (2015) ChrF: character *n*-gram F-score for automatic MT evaluation. In: Proceedings of the tenth workshop on Statistical Machine Translation (WMT2015), Lisbon, pp 392–395

Popović M (2017) Comparing language related issues for NMT and PBMT between German and English. Prague Bull Math Linguist 108(1):209–220

Popović M, Arcan M (2015) Identifying main obstacles for statistical machine translation of morphologically rich South Slavic languages. In: The 18th annual conference of the European Association for Machine Translation (EAMT 2015), Antalya, pp 97–104

Popović M, Arcan M (2016) PE2rr corpus: manual error annotation of automatically pre-annotated MT post-edits. In: Proceedings of the tenth international conference on Language Resources and Evaluation (LREC 2016)

Popović M, Ney H (2006) Error analysis of verb inflections in Spanish translation output. In: Proceedings of the TC-Star workshop on speech-to-speech translation, Barcelona, pp 99–103

Popović M, Ney H (2007) Word error rates: decomposition over POS classes and applications for error analysis. In: Proceedings of the 2nd workshop on Statistical Machine Translation (WMT 2007), Prague, pp 48–55

Popović M, Ney H (2011) Towards automatic error analysis of machine translation output. Comput Linguist 37(4):657–688

Popović M, de Gispert A, Gupta D, Lambert P, Ney H, Mariño JB, Federico M, Banchs R (2006) Morpho-syntactic information for automatic error analysis of statistical machine translation output. In: Proceedings on the 1st workshop on Statistical Machine Translation, New York, pp 1–6

Popović M, Lommel A, Burchardt A, Avramidis E, Uszkoreit H (2014) Relations between different types of post-editing operations, cognitive effort and temporal effort. In: Proceedings of the 7th annual conference of the European Association for Machine Translation (EAMT 2014), pp 191–198

Popović M, Arcan M, Avramidis E, Burchardt A, Lommel A (2015) Poor man's lemmatisation for automatic error classification. In: The 18th annual conference of the European Association for Machine Translation (EAMT 2015), pp 105–112

Popović M, Arcan M, Klubicka F (2016) Language related issues for machine translation between closely related South Slavic languages. In: Proceedings of the third workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2016), Osaka, pp 43–52

Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J (2006) A study of translation edit rate with targeted human annotation. In: Proceedings of AMTA 2006, the 7th conference of the Association for Machine Translation in the Americas, Cambridge, pp 223–231

Stymne S (2011) Blast: a tool for error analysis of machine translation output. In: Proceedings of the 49th annual meeting of the Association for Computational Linguistics – Human Language Technologies (HLT 2011): Systems Demonstrations, Portland, pp 56–61

Stymne S, Ahrenberg L (2012) On the practice of error analysis for machine translation evaluation. In: Proceedings of the 8th international conference on Language Resources and Evaluation (LREC 2012), Istanbul

Toral A, Sánchez-Cartagena VM (2017) A multifaceted evaluation of neural versus statistical machine translation for 9 language directions. In: Proceedings of the 15th conference of the European chapter of the Association for Computational Linguistics (EACL 2017), Valencia

Toral A, Naskar SK, Gaspari F, Groves D (2012) DELiC4MT: a tool for diagnostic MT evaluation over user-defined linguistic phenomena. Prague Bull Math Linguist 98:121–132

Vilar D, Xu J, D'Haro LF, Ney H (2006) Error analysis of statistical machine translation output. In: Proceedings of 5th international conference on Language Resources and Evaluation (LREC 2006), Genoa, pp 697–702

Vogel S, Ney H, Tillmann C (1996) HMM-based word alignment in statistical translation. In: Proceedings of the 16nd international conference on Computational Linguistics (CoLing 1996), Copenhagen, Denmark, pp 836–841

Vossen P, Rigau G, Agirre E, Soroa A, Monachini M, Bartolini R (2010) KYOTO: an open platform for mining facts. In: Proceedings of the 6th workshop on Ontologies and Lexical Resources (Ontolex 2010), Beijing, pp 1–10

Wang B, Zhou M, Liu S, Li M, Zhang D (2014) Woodpecker: an automatic methodology for machine translation diagnosis with rich linguistic knowledge. J Inf Sci Eng 30(5):1407–1424

Zaretskaya A, Vela M, Pastor GC, Seghiri M (2016) Measuring post-editing time and effort for different types of machine translation errors. New Voice Trans Stud 15:63–92

Zeman D, Fishel M, Berka J, Bojar O (2011) Addicter: what is wrong with my translations? Prague Bull Math Linguist 96:79–88

Zhou M, Wang B, Liu S, Li M, Zhang D, Zhao T (2008) Diagnostic Evaluation of machine translation systems using automatically constructed linguistic check-points. In: Proceedings of the 22nd international conference on Computational Linguistics (CoLing 2008), Manchester, pp 1121–1128