From Human to Automatic Error Classification for Machine Translation Output

Maja Popović and Aljoscha Burchardt
German Research Center for Artificial Intelligence (DFKI)
Language Technology Group (LT)
Berlin, Germany
{maja.popovic,aljoscha.burchardt}@dfki.de

Abstract

Future improvement of machine translation systems requires reliable automatic evaluation and error classification measures to avoid time and money consuming human classification. In this article, we propose a new method for automatic error classification and systematically compare its results to those obtained by humans. We show that the proposed automatic measures correlate well with human judgments across different error classes as well as across different translation outputs on four out of five commonly used error classes.

1 Introduction and related work

The evaluation of machine translation output is an intrinsically difficult task. Human evaluation is expensive and time consuming. Therefore a great deal of effort has been spent on finding measures that correlate well with human judgements when ranking translation systems for quality (see for example (Callison-Burch et al., 2009; Callison-Burch et al., 2010)). A considerable amount of work has been put into the improvement of these measures. However, most of the work has been focused just on ranking between different machine translation systems. While ranking systems is an important first step towards their improvement, it does not provide enough scientific insights. Researchers often would find it helpful to get answers to questions such as *What is a particular strength/weakness of my system? What kind of errors does the system make most often? Does a particular modification improve some aspect of the system, even if it does not improve the overall score? Does a worse–ranked system outperform a better–ranked one in any aspect?*, etc.

In order to answer such questions, a framework for human error analysis and error classification has been proposed in (Vilar et al., 2006), where a classification scheme based on (Llitjós et al., 2005) is presented together with a detailed analysis of the obtained results. The method has become widely used in recent years (Avramidis and Koehn, 2008; Max et al., 2008; Khalilov and Fonollosa, 2009; Li et al., 2009). Still, human error classification is resource-intensive and might become practically unfeasible when translating into many languages.

As for automatic methods, an approach for automatic identification of patterns in translation output using POS sequences is proposed in (Lopez and Resnik, 2005) in order to see how well a translation system is capable of capturing systematic reordering patterns. Using relative differences between Word Error Rate (WER) and Position-independent Word Error Rate (PER) for nouns, adjectives and verbs has been proposed in (Popović et al., 2006) for the estimation of inflectional and reordering errors. A method based on WER and PER decomposition for discovering inflectional errors and missing words is presented in (Popović and Ney, 2007). Zhou (2008) proposed a diagnostic evaluation of linguistic check-points obtained automatically by aligning parsed source and target sentences. However, to our best knowledge, there has been no attempt to design a set of automatic metrics which covers the error categories from (Vilar et al., 2006) in a systematic manner.

In this work, we first define five error categories based on those described in (Vilar et al., 2006) and present the results for these categories obtained by human evaluators and by a novel automatic tool

based on the method proposed in (Popović and Ney, 2007). We calculate correlations between human and automatic error classification results, both across different error classes as well as across different translation outputs. Finally, we perform a deep analysis of the obtained results in order to better understand the differences between human and automatic evaluation.

2 Error classification

The two main goals of the proposed automatic method for error analysis and classification are to be able:

- to estimate the distribution of errors over the error classes in order to determine which error types are particularly problematic for a given translation system;
- to estimate the differences between the numbers of errors in each class for different translation outputs in order to compare translation systems.

The starting point for the automatic error classification proposed in this work is the identification of actual words contributing to the Word Error Rate (WER) (Levenshtein, 1966) and to the recall- and precision-based Position-independent Error Rates called Reference PER (RPER) and Hypothesis PER (HPER) (Popović and Ney, 2007). The WER errors are marked as substitutions, deletions or insertions. The RPER errors represent the words in the reference which do not appear in the hypothesis, and the HPER errors the words in the hypothesis which do not appear in the reference. If multiple reference translations are available, the reference with the lowest WER score is choosen for all metrics.

Once these words have been identified, the following error categories based on the classification scheme used in (Vilar et al., 2006) are defined:

- inflectional errors — an inflectional error occurs if the base form of the generated word is correct but the full form is not.
- reordering errors — a word which occurs both in the reference and in the hypothesis thus not contributing to RPER or HPER but is marked as a WER error is considered as a reordering error.
- missing words — a word which occurs as deletion in WER errors and at the same time occurs as RPER error without sharing the base form with any hypothesis error is considered as missing.
- extra words — a word which occurs as insertion in WER errors and at the same time occurs as HPER error without sharing the base form with any reference error is considered as extra.
- incorrect lexical choice — a word which belongs neither to inflectional errors nor to missing or extra words is considered as lexical error.

The presented method is language-independent, however availability of base forms for the particular target language is a requisite.

Human error classification

As there are often several correct translations of a given source sentence that correspond more or less to the given reference translation(s), human error analysis can be carried out in various ways. Errors can be counted by doing a direct strict comparison between the given reference and the translation outputs, but much more flexibility can be allowed: substitution of words and expresions by synonyms, syntactically correct different word order, etc, which is a more natural way. It is also possible to use the references only for the semantic aspect, i.e. to look only whether the main meaning is preserved. It is even possible not to use a reference translation at all, but compare the translation output with the source text.

The human error classification is definitely not unambigous — often it is not easy to determine in which particular error category some error exactly belongs, sometimes one word can be assigned to more than one category, and variations between different human evaluators are possible. Especially difficult is disambiguating between incorrect lexical choice and missing words or extra words. Furthermore, a choice of words to be assigned to reordering class may vary. Some typical examples are shown in Table 1. In the first example, one possible interpretation is that `All-People Headquarters` are missing words and `Pan Country` are extra words. However, it could also be considered that all words represent incorrect lexical choice. In addition, `in`

reference translation	obtained output	error classes
in the General Assembly resolution, All-People Headquarters said ...	Pan Country in the General Assembly resolution, said ...	missing+extra or lexical?
a more serious problem ...	a problem more serious ...	reordering errors?

Table 1: Examples of ambiguous error classficiation.

`the General Assembly` may or may not be considered as reordering error. The second example presents a typical example of reordering ambiguity: which words should be assigned to this class: `more serious`, or `problem`, or all of them? Despite of these ambiguities, such scheme for error classification has proven to be useful and an automatisation of the process is needed. More elaborate classification schemes using sets of errors per word are left for future work.

In this work, two types of human error analysis using reference translations are carried out in order to make a fair comparison with the automatic method: a strict one (comparing with a reference) and a flexible one (syntactically correct differences and word order and substitutions by synonyms are not considered as errors). The flexible type of error analysis identifies much fewer words as errors.

3 Experimental results

3.1 Experimental set-up

For the human and automatic error classifications described in the previous sections, we used six English translation outputs obtained by state-of-the-art statistical phrase-based translation systems in the framework of the GALE[1] project and the fourth Workshop on Statistical Machine Translation[2] (WMT). Two GALE outputs are translations from Arabic into English, and the third is a result of Chinese–to–English translation. All three WMT outputs are translations from the same German text into English, thus being appropriate for comparison of different translation systems. For each translation output, only one reference translation was available. For the GALE texts, the strict human error analysis is carried out, and for the WMT texts the flexible one. TreeTagger[3] was used for obtaining the base forms of the words for the automatic error classification.

[1]GALE – Global Autonomous Language Exploitation. http://www.arpa.mil/ipto/programs/gale/index.htm

[2]Fourth Workshop on Statistical Machine Translation. http://www.statmt.org/wmt09/

[3]http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

3.2 Results and correlations

The results of both human and automatic error analysis for all texts and all error classes are presented in Table 2 in the form of raw error counts. In addition, the Pearson (r) and Spearman rank (ρ) correlation coefficients between human and automatic results are calculated for each translation output across error classes (rightmost column). Since all WMT outputs are translations of the same source text, the correlations are presented also for each error class across different translation outputs (last row). These correlation coefficients are very high, both across the error classes and across the translation outputs. For the WMT outputs, the correlations across the error classes are slightly lower than for the GALE outputs; this could be expected due to the more flexible criteria for the human error classification. In addition, it can be noted that the extra words category has the weakest correlation across different translation outputs.

The results show that the automatic method can successfully substitute human analysis in order to answer the questions that the overall ranking evaluation metrics cannot. The automatic method is well capable of detecting weak and strong points of particular translation system as well as of comparing different translation systems. Nevertheless, there are certain differences between human and automatic classification, and it would be useful to better understand them: Are all errors detected by humans successfully covered by the automatic method? Why does the automatic tool assign much more reordering errors than human evaluators? How does the automatic method cope with disambiguation between lexical errors and missing/extra words? Why are the correlations for extra words lower than others?

3.3 Analysis of the differences

In order to answer the above questions, recall and precision of all error classes are presented in the form of a confusion matrix. Recall shows how many errors classified by humans are successfully

(a) GALE translation outputs

GALE (BLEU)	inflection	order	missing	extra	lexical	ρ	r
ArEn1 (59.7%)	20/23	39/66	79/63	127/137	135/147	0.90	0.96
ArEn2 (72.1%)	22/24	30/41	97/102	73/76	140/131	1.00	0.99
CnEn (58.0%)	38/40	127/171	288/244	95/117	203/239	1.00	0.93

(b) WMT translation outputs

WMT (BLEU)	inflection	order	missing	extra	lexical	ρ	r
DeEn1 (16.9%)	12/32	60/235	204/199	52/40	189/521	0.70	0.72
DeEn2 (18.4%)	16/44	41/212	172/200	30/56	163/495	0.7	0.74
DeEn3 (17.2%)	17/46	100/274	107/153	68/99	171/508	0.90	0.91
ρ	1.00	1.00	0.60	0.5	1.00		
r	0.90	0.99	0.90	0.62	0.96		

Table 2: Raw error counts N_{hum}/N_{aut} obained by human (left) and automatic (right) error analysis for the GALE (a) and for the WMT (b) translation outputs; Spearman (left) and Pearson (right) correlation coefficients for each translation output across error categories (last two columns) and for each error category across different WMT translation outputs (last two rows). For each translation output, the BLEU score is given as illustration.

covered by the automatic tool, and precisions how many automatically classified errors are correct, i.e. assigned to the same class by humans. The results are presented for one translation output of each set, namely for the GALE output ArEn1 and for the WMT output DeEn1.

3.3.1 GALE translation outputs

Table 3 shows the results for the GALE ArEn1 output. For each error class, both recall and precision are high – errors detected by humans are sucesfully detected by the automatic tool too, and at the same time errors detected by the automatic method are marked as errors by humans as well. However, the precision of reordering errors is lower than for other categories – about a third of automatically detected reordering errors are not considered as erroneous words by humans. Further inspection showed that the majority of such words are articles, punctuations, the conjunctions "and" and "or" as well as some prepositions, i.e. words which occur frequently. Since they often appear several times in one sentence, the automatic tool does not see them as RPER/HPER errors, but depending on their position they are often marked as WER errors and thus classified as reordering errors. A similar phenomenon also leads to a number of extra words and lexical errors in the hypothesis which are not detected as errors by the automatic tool. And if there were more frequent words in the reference than in the hypothesis, the same could happen with missing words and reference lexical errors. Introducing some kind of information about the word position into the process can diminish these discrepances. Apart from that, a certain number of reordering errors is distributed differently over words as explained in Table 1 thus leading to confusions "x-reord" and "reord-x". As for disambiguation of lexical errors vs missing/extra words, both recall and precision confusions can be observed, though lexical errors have a rather high recall. This means that lexical errors detected by humans are very well covered by the automatic tool, but a number of human annotated extra and missing words are also considered as lexical errors.

Examples of human and automatic error analysis are presented in Table 4. The first sentence illustrates a total agreement between the human and automatic error classification. In the second sentence, the words `Japanese` and `friendly` are classified into the same category both by the human and by the automatic analysis. The words `feeling for` represent an example where the human analysis assigns the error to the missing words category, but the automatic analysis classifies it as a lexical error. Similarly, the words `can feel` are considered as extra words by humans, but as lexical errors by the automatic tool.

3.3.2 WMT translation outputs

The results for the WMT DeEn1 output are presented in Table 5. The main difference in compar-

(a) Reference translation for the GALE ArEn1 output.

ArEn1 ref	inflection	order	missing	lexical	x
inflection	**78.9/78.9**	/	2.2/10.5	0.8/5.3	0.1/5.3
order	/	**92.5/51.4**	8.8/11.1	3.2/5.6	1.8/31.9
missing	/	/	**53.8/81.7**	4.8/10.0	0.4/8.3
lexical	15.8/2.1	2.5/0.7	29.7/19.3	**85.5/75.7**	0.2/2.1
x	5.3/0.1	5.0/0.2	5.5/0.4	5.6/0.5	**97.5/98.8**

(b) ArEn1 hypothesis translation.

ArEn1 hyp	inflection	order	extra	lexical	x
inflection	**81.0/89.5**	/	0.7/5.3	/	0.1/5.3
order	4.8/1.4	**90.2/51.4**	3.6/6.9	5.8/12.5	1.5/27.8
extra	4.8/1.0	/	**53.3/72.3**	15.4/23.8	0.2/3.0
lexical	4.8/0.8	2.4/0.8	15.3/15.9	**64.1/75.8**	0.8/6.8
x	4.8/0.1	7.3/0.2	27.0/2.8	14.7/1.7	**97.5/95.2**

Table 3: Recall (left) and precision (right) values for the GALE ArEn1 translation output:
(a) reference translation, (b) hypothesis. The columns represent the error classes obtained by human evaluators, the rows represent the classes obtained automatically. The class "x" stands for "no error detected".

reference:	... of local party committees . ***Secretaries*** of the Commission ...
hypothesis:	... of local party committees of the ***provincial*** Commission ...
errors:	Secretaries – missing(hum,aut)
	provincial – extra(hum,aut)
reference:	... , although the ***Japanese friendly*** **feelings for** China added an increase , ...
hypothesis:	... , although China **can feel** the ***Japanese*** increase , ...
errors:	Japanese – order(hum,aut)
	friendly – missing(hum,aut)
	feelings for – missing(hum)/lexical(aut)
	can feel – extra(hum)/lexical(aut)

Table 4: Examples of human and automatic error analysis from the GALE translation outputs: words in bold italic are assigned to the same error category both by human and automatic error analysis, and words only in bold represent differences.

ison with the GALE results is that the precisions of all error classes are much lower – the automatic tool identifies much more errors than human evaluators. This is especially notable for reordering and for lexical errors – the reason is the flexible human evaluation which allows synonyms and different word orders. Nevertheless, the recall values are very high (except for extra words), meaning that the automatic tool is capable of discovering errors detected by human evaluators also when the flexible (more natural) human classification is carried out.

The high number of reordering errors is again mostly due to the frequent words, but there is also a number of other words as well since the flexible human classification allows more word orders. In addition, identifying different words as reordering errors happens more often. There is also a number of reordering errors which the automatic method considers as lexical errors: the reason for that are the synonyms or different expressions. A different way of expression is also the reason for the higher number of automatically detected inflectional errors, for example `patients' health -- health of the patient` or `is building -- builds`.

For this set, the confusion between lexical errors vs missing/extra words is also present, especially for the extra words – the major part of extra

(a) Reference translation for the WMT DeEn1 output.

DeEn1 ref	inflection	order	missing	lexical	x
inflection	**92.3**/37.5	1.6/3.1	2.0/12.5	1.6/9.4	1.1/37.5
order	/	**61.3**/15.3	5.9/4.8	2.6/2.0	17.3/**77.8**
missing	/	6.5/2.1	45.8/48.4	16.6/16.7	5.7/32.8
lexical	7.7/0.2	11.3/1.4	42.9/17.5	**78.2**/30.3	22.6/**50.6**
x	/	19.4/1.9	3.4/1.1	1.0/0.3	**53.4/96.6**

(b) DeEn1 hypothesis translation.

DeEn1 hyp	inflection	order	extra	lexical	x
inflection	**92.3**/37.5	5.4/12.5	/	2.6/12.5	1.1/37.5
order	/	**51.4**/15.3	14.8/3.2	4.5/2.8	17.8/**78.6**
extra	/	1.4/3.2	16.7/29.0	3.2/16.1	1.5/**51.6**
lexical	7.7/0.2	24.3/4.0	**57.4**/6.9	**85.8**/29.6	24.4/**59.3**
x	/	17.6/2.1	11.1/1.0	3.9/1.0	**55.3/96.0**

Table 5: Recall (left) and precision (right) values for the WMT DeEn1 translation output: (a) reference translation, (b) hypothesis. The columns represent the error classes obtained by human evaluators, the rows represent the classes obtained automatically. The class "x" stands for "no error detected".

words recall is actually confusion with lexical errors. There is also a number of extra words which are assigned to reordering errors or correct words – these are again mostly frequent words. These are the reasons why extra words have the lowest correlation coefficients across the translation outputs – this error category is not particularly reliable for comparing different translation systems. Lexical errors on the other hand have very high recall – as for the GALE task, those detected by humans are successfully covered by the automatic tool. However, because of the synonyms, the precision is low – the major part of the automatically detected lexical errors are actually correct words. Using synonym lists can increase this precision and also decrease the number of reordering errors classified as lexical errors.

Table 6 presents examples of human and automatic error analysis for the WMT data. The first example illustrates total agreement. In addition, it also illustrates a case where the reordering errors could be defined in a different way, both by human evaluators and by the automatic tool: the word group `coffee and newspapers` could be considered as reordering error. This phenomenon can also be seen in the second sentence, namely `famous journalist Gustav Chalupa` may also be considered as a reordering error. Furthermore, this sentence illustrates confusions between lexical errors and missing/extra words, as well as why the number of the lexical errors is significantly higher for the automatic tool. The tool considers `this`, `the` and `Lamborgini` as lexical errors, as well as `born in České Budějovice/from Budweis`. However, the humans considered the first three words as missing or extra words, and the rest beeing synonyms is not considered as error at all — Budweis is English name for the Czech town České Budějovice.

4 Conclusions and outlook

In this work we have proposed a systematic method for automatic error classification of machine translation output. The method detects five error classes commonly used in human error analysis: inflectional errors, reordering errors, missing words, extra words and incorrect lexical choice. We have shown that the error classification results obtained by this approach correlate very well with the results of human error analysis with Spearman and Pearson correlation coefficients over 0.7 and mostly around 0.9, both across different error categories within one translation output as well as across different translation outputs within one error category. The automatic metrics also have high recall, i.e. the method is well capable of finding the errors detected by human evaluators. Hence, the presented automatic method can successfully replace human error analysis in order to get bet-

reference:	***Passengers can get*** coffee and newspapers ***when*** boarding .
hypothesis:	Coffee and newspapers ***can passengers in*** boarding .
errors:	Passengers can – order(hum,aut)
	get – missing(hum,aut)
	when – lexical(hum,aut)
	in – lexical(hum,aut)
reference:	The **famous journalist Gustav Chalupa** , **born in České Budějovice** , ***also confirms*** **this** .
hypothesis:	The ***also confirms*** **the famous** ***Austrian*** **journalist Gustav Chalupa** , **from Budweis Lamborghini** .
errors:	famous journalist Gustav Chalupa – order(aut)
	born in České Budějovice – lex(aut)
	also confirms – order(hum,aut)
	this – missing(hum)/lexical(aut)
	the – extra(hum)/lexical(aut)
	Austrian – extra(hum,aut)
	from Budweis – lexical(aut)
	Lamborghini – extra(hum)/lexical(aut)

Table 6: Examples of human and automatic error analysis from the WMT translation outputs: words in bold italic are assigned to the same error category both by human and automatic error analysis, and words only in bold represent differences.

ter insight about the strengths and weaknesses of one translation system as well as about the differences between various translation systems. Only the extra word class has proven not to be stable and reliable enough.

In a detailed qualitative analysis of typical problems related to both human and automatic classification of translation errors, we pointed out promising future work. Introducing synonym lists and information about word positions can further help the presented automatic method to increase precision, as could going from word to phrase level. Other directions that would address the intrinsic difficulty of the error classification tasks are adding probabilities to error classes and allowing the assignment of multiple errors per word.

The method is currently being tested and further developed in the framework of the TARAXÜ project[4]. In this project, three industry and one research partner aim to develop a hybrid machine translation architecture that satisfies current industry needs, which includes a number of large-scale evaluation rounds involving various target languages: English, French, German, Czech, Spanish, Russian, Chinese and Japanese.

[4]http://taraxu.dfki.de/

Acknowledgments

This work has partly been developed within the TARAXÜ project financed by TSB Technologiestiftung Berlin – Zukunftsfonds Berlin, co-financed by the European Union – European fund for regional development. Special thanks to David Vilar and Eleftherios Avramidis.

References

Avramidis, Eleftherios and Philipp Koehn. 2008. Enriching morphologically poor languages for statistical machine translation. In *Proceedings of the 46rd Annual Meeting of the Association for Computational Linguistics (ACL 08)*, pages 763–770, Columbus, Ohio, June.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece, March.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR (WMT 10)*, pages 17–53, Uppsala, Sweden, July.

Khalilov, Maxim and José A. R. Fonollosa. 2009. N-gram-based statistical machine translation versus syntax augmented machine translation: comparison and system combination. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computation Linguistic (EACL 09)*, pages 424–432, Athens, Greece, March.

Levenshtein, Vladimir Iosifovich. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10(8):707–710, February.

Li, Jin-Ji, Jungi Kim, Dong-Il Kim, and Jong-Hyeok Lee. 2009. Chinese syntactic reordering for adequate generation of korean verbal phrases in chinese-to-korean smt. In *Proceedings of the Fourth Workshop on Statistical Machine Translation (WMT 09)*, pages 190–196, Athens, Greece, March.

Llitjós, Ariadna Font, Jaime G. Carbonell, and Alon Lavie. 2005. A framework for interactive and automatic refinement of transfer-based machine translation. In *Proceedings of the 10th Annual Conference of the European Association for Machine Translation (EAMT 05)*, pages 87–96, Budapest, Hungary, May.

Lopez, Adam and Philip Resnik. 2005. Pattern visualization for machine translation output. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, pages 12–13, Vancouver, Canada, October.

Max, Aurélien, Rafik Makhloufi, and Philippe Langlais. 2008. Explorations in using grammatical dependencies for contextual phrase translation disambiguation. In *Proceedings of the 12th Annual Conference of the European Association for Machine Translation (EAMT 08)*, pages 114–119, Hamburg, Germany, September.

Popović, Maja and Hermann Ney. 2007. Word Error Rates: Decomposition over POS classes and Applications for Error Analysis. In *Proceedings of the 2nd ACL 07 Workshop on Statistical Machine Translation (WMT 07)*, pages 48–55, Prague, Czech Republic, June.

Popović, Maja, Adrià de Gispert, Deepa Gupta, Patrik Lambert, Hermann Ney, José B. Mariño, Marcello Federico, and Rafael Banchs. 2006. Morpho-syntactic Information for Automatic Error Analysis of Statistical Machine Translation Output. In *Proceedings of the 1st NAACL 06 Workshop on Statistical Machine Translation (WMT 06)*, pages 1–6, New York, NY, June.

Vilar, David, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error Analysis of Statistical Machine Translation Output. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 06)*, pages 697–702, Genoa, Italy, May.

Zhou, Ming, Bo Wang, Shujie Liu, Mu Li, Dongdong Zhang, and Tiejun Zhao. 2008. Diagnostic evaluation of machine translation systems using automatically constructed linguistic check-points. *Proceedings of the 22nd International Conference on Computational Linguistics (CoLing 2008)*, pages 1121–1128, August.