Class error rates for evaluation of machine translation output

Maja Popović
German Research Center for Artificial Intelligence (DFKI)
Language Technology (LT), Berlin, Germany
maja.popovic@dfki.de

Abstract

We investigate the use of error classification results for automatic evaluation of machine translation output. Five basic error classes are taken into account: morphological errors, syntactic (reordering) errors, missing words, extra words and lexical errors. In addition, linear combinations of these categories are investigated. Correlations between the class error rates and human judgments are calculated on the data of the third, fourth, fifth and sixth shared tasks of the Statistical Machine Translation Workshop. Machine translation outputs in five different European languages are used: English, Spanish, French, German and Czech. The results show that the following combinations are the most promising: the sum of all class error rates, the weighted sum optimised for translation into English and the weighted sum optimised for translation from English.

1 Introduction

Recent investigations have shown that it is possible to carry out a reliable automatic error analysis of a given translation output in order to get more information about actual errors and details about particular strengthnesses and weaknesses of a systeml (Popović and Ney, 2011). The obtained results correlate very well with the human error classification results. The question we try to answer is: how the class error rates correlate with the human evaluation (ranking) results? As a first step, we investigate the correlations of five basic class error rates with human rankings. In the next step, linear combinations (sums) of basic class error rates are investigated.

Spearman's rank correlation coefficients on the document (system) level between all the metrics and the human ranking are computed on the English, French, Spanish, German and Czech texts generated by various translation systems in the framework of the third (Callison-Burch et al., 2008), fourth (Callison-Burch et al., 2009), fifth (Callison-Burch et al., 2010) and sixth (Callison-Burch et al., 2011) shared translation tasks.

2 Class error rates

In this work, the method proposed in (Popović and Ney, 2011) is used, i.e. classification of the translation errors into five basic categories based on the Word Error Rate (WER) (Levenshtein, 1966) together with the recall- and precision-based Position-independent Error Rates called Reference PER (RPER) and Hypothesis PER (HPER).

As a result of an error classification, two values are usually of interest: raw error counts for each error class, and error rates for each class, i.e. raw error counts normalised over the total number of running words. Which of the values is preferred depends of the exact task. For example, if only a distribution of error types within a translation output is of interest, the raw error counts are sufficient. On the other hand, if we want to compare different translation outputs, normalised values i.e. error rates are more suitable. Therefore they are appropriate candidates to be used for the evaluation task.

In this work, we explore the error rates calculated on the word level as well as on the block level, where

a group of consecutive words labelled with the same error category is called a block. The normalisation in both cases is carried out over the total number of running words. Therefore the block level error rate for a particular error class is always less or equal than the corresponding word level error rate.

2.1 Basic class error rates

The following five basic class error rates are explored:

INFER (**inf**lectional **e**rror **r**ate):
Number of words translated into correct base form but into incorrect full form, normalised over the hypothesis length.

RER (**r**eordering **e**rror **r**ate):
Number of incorrectly positioned words normalised over the hypothesis length.

MISER (**mis**sing word **e**rror **r**ate):
Number of words which should appear in the translation hypothesis but do not, normalised over the reference length.

EXTER (**ext**ra word **e**rror **r**ate):
Number of words which appear in the translation hypothesis but should not, normalised over the hypothesis length.

LEXER (**lex**ical **e**rror **r**ate):
Number of words translated into an incorrect lexical choice in the target language (false disambiguation, unknown/untranslated word, incorrect terminology, etc.) normalised over the hypothesis length.

Table 1 presents an example of word and block level class error rates. Each erroneous word is labelled with the corresponding error category, and the blocks are marked within the parentheses { and }. The error rates on the block level are marked with a letter "b" at the beginning. It should be noted that the used method at its current stage does not enable assigning multiple error tags to one word.

2.2 Combined error rates (sums)

The following linear combinations (sums) of the basic class error rates are investigated:

reference:
The famous journalist Gustav Chalupa , born in České Budějovice , also confirms this .
hypothesis containing 14 running words:
The also confirms the famous Austrian journalist Gustav Chalupa , from Budweis Lamborghini .
hypothesis labelled with error classes:
The $\{also_{order}\ confirms_{order}\}$ $\{the_{extra}\}\ \{famous_{order}\}\ \{Austrian_{extra}\}$ $\{journalist_{order}\ Gustav_{order}\ Chalupa_{order}\}$, $\{from_{lex}\ Budweis_{lex}\ Lamborghini_{lex}\}$.
class error rates:
word order: RER = 6/14 = 42.8% bRER = 3/14 = 21.4%
extra words: EXTER = 2/14 = 14.3% bEXTER = 2/14 = 14.3%
lexical errors: LEXER = 3/14 = 21.4% bLEXER = 1/14 = 7.1%

Table 1: Example of word and block level class error rates: the word groups within the parentheses { and } are considered as blocks; all error rates are normalised over the hypothesis length, i.e. 14 running words.

WΣER (**sum** of **w**ord level **e**rror **r**ates)[1] :
Sum of all basic class error rates on the word level;

BΣER (**sum** of **b**lock level **e**rror **r**ates):
Sum of all basic class error rates on the block level;

WBΣER (**sum** of **w**ord and **b**lock level **e**rror **r**ates):
Arithmetic mean of WΣER and BΣER.

[1]This error rate has already been introduced in (Popović and Ney, 2011) and called ΣER; however, for the sake of clarity, in this work we will call it WΣER, i.e. word level ΣER.

XEnΣER (**X**→**En**glish **sum** of **e**rror **r**ates):
Linear interpolation of word level and block level class error rates optimised for translation into English;

EnXΣER (**En**glish→**X sum** of **e**rror **r**ates):
Linear interpolation of word level and block level class error rates optimised for translation from English.

For the example sentence shown in Table 1, WΣER = 84.7%, BΣER = 46.2% and WBΣER = 65.4%. XEnΣER and EnXΣER are weighted sums which will be explained in the next section.

The prerequisite for the use of the described metrics is availability of an appropriate morphological analyser for the target language which provides base forms of the words.

3 Experiments on WMT 2008, 2009, 2010 and 2011 test data

3.1 Experimental set-up

The class error rates described in Section 2 were produced for outputs of translations from Spanish, French, German and Czech into English and vice versa using Hjerson (Popović, 2011), an open-source tool for automatic error classification. Spanish, French, German and English base forms were produced using the TreeTagger[2], and the Czech base forms using Morče (Spoustová et al., 2007). In this way, all references and hypotheses were provided with the base forms of the words.

For each error rate, the system level Spearman correlation coefficients ρ with human ranking were calculated for each document. In total, 40 correlation coefficients were obtained for each error rate – twelve English outputs from the WMT 2011, 2010 and 2009 task and eight from the WMT 2008 task, together with twenty outputs in other four target languages. For further analysis, the obtained correlation results were summarised into the following three values:

- *mean*
  average correlation coefficient;
- *rank*$>$
  percentage of documents where the particular error rate has better correlation than the other error rates;
- *rank*$\geq$
  percentage of documents where the particular error rate has better or equal correlation than the other error rates.

3.2 Comparison of basic class error rates

Our first experiment was to compare correlations for the basic set of class error rates in order to investigate a general behaviour of each class error rate and to see if some of the error categories are particularly (in)convenient for the evaluation task. Since certain differences between English and non-English translation outputs are observed for some error classes, the values described in the previous section were also calculated separately.

Table 2 presents the results of this experiment. The *mean* values over all documents, over the English documents and over the non-English documents are shown.

According to the overall *mean* values, the most promising error categories are lexical and reordering errors. However, the *mean* values for English outputs are significantly different than those for non-English outputs: the best error classes for English are in deed lexical and reordering errors, however for the non-English outputs the inflectional errors and missing words have higher correlations. On the other hand, for the English outputs missing words have even negative correlations, whereas correlations for inflectional errors are relatively low. The extra word class seems to be the least convenient in general, especially for non-English outputs.

Therefore, the *rank*$\geq$ values were calculated only separately for English and non-English outputs, and the previous observations were confirmed: for the English outputs lexical and reordering errors are the most relevant, whereas for the non-English outputs all classes except extra words are almost equally important.

Apart from this, it can be noticed that the grouping of words into blocks significantly improves correlation for reordering errors. The reason for this is ambiguity of tagging words as reordering errors.

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

error rate	*mean*			*rank*$\geq$	
	overall	x→en	en→x	x→en	en→x
INFER	0.398	0.190	*0.595*	46.2	*71.7*
RER	0.360	0.344	0.373	*53.8*	*51.1*
MISER	0.173	-0.101	0.434	26.3	**54.4**
EXTER	0.032	0.212	-0.195	42.7	12.2
LEXER	0.508	**0.669**	0.355	**86.0**	*58.3*
bINFER	0.423	0.211	**0.624**	47.9	**75.6**
bRER	0.508	**0.594**	0.426	**78.3**	**60.0**
bMISER	0.169	-0.121	**0.446**	21.1	*53.9*
bEXTER	-0.031	0.186	-0.238	36.8	10.0
bLEXER	0.515	*0.634*	0.402	*79.5*	**62.8**

Table 2: *mean* and *rank*$\geq$ values for each basic word level and block level error rate over all documents, over English documents and over non-English documents.

For example, if the translation reference is "a very good translation", and the obtained hypothesis is "a translation very good" , one possibility is to mark the word "translation" as reordering error, another possibility is to mark the words "very good" as reordering errors, and it is also possible to mark all the words as reordering errors. In such cases, the grouping of consecutive word level errors into blocks is beneficial.

3.3 Comparison of error rate sums

A first step towards combining the basic class error rates was investigation of simple sums, i.e. WΣER, BΣER as well as WBΣER as arithmetic mean of previous two. The overall average correlation coefficients of the sums were shown to be higher than those of the basic class error rates. Further experiments have been carried out taking into account the results described in the previous section. Firstly, extra word class was removed from all sums, however no improvement of correlation coefficients was observed. Then the sums containing only the most promising error categories separately for English and non-English output were investigated, but this also resulted in no improvements. Finally, we introduced weights for each translation direction according to the *rank*$\geq$ value for each of the basic class error rates (see Table 2), and this approach was promising. In this way, the specialised sums XENΣER and ENXΣER were introduced.

In Table 3 the results for all five error rate sums are presented. *mean*, *rank*$>$ and *rank*$\geq$ values are presented over all translation outputs, over English outputs and over non-English outputs. As already mentioned, the overall correlation coefficients of the sums are higher than those of the basic class error rates. This could be expected, since summing class error rates is oriented towards the overall quality of the translation output whereas the class error rates are giving more information about details.

According to the overall values, the best error rate is combination of all word and block level class error rates, i.e. WBΣER followed by the block sum BΣER, whereas the WΣER and the specialised sums XENΣER and ENXΣER have lower correlations. For the translation into English, this error rate is also very promising, followed by the specialised sum XENΣER. On the other hand, for the translation from English, the most promising error rates are the block sum BΣER and the corresponding specialised sum ENXΣER. Following these observations, we decided to submit WBΣER scores for all translation outputs together with XENΣER and ENXΣER scores, each one for the corresponding translation direction. In addition, we submitted BΣER scores since this error rate also showed rather good results, especially for the translation out of English.

4 Conclusions

The presented results show that the error classification results can be used for evaluation and ranking of machine translation outputs. The most promising way to do it is to sum all word level and block level error rates, i.e. to produce the WBΣER error rate. This error rate has eventually been submitted to the WMT 2012 evaluation task. In addition, the next best metrics have been submitted, i.e. the block level sum BΣER for all translation directions, and the specialised sums XENΣER and ENXΣER each for the corresponding translation outputs.

The experiments described in this work are still at early stage: promising directions for future work are better optimisation of weights[3], further investigation of each language pair and also of each non-English

[3]First steps have already been made in this direction using an SVM classifier, and the resulting evaluation metric has also been submitted to the WMT 2012.

error rate	*mean*			*rank*$\geq$			*rank*$>$		
	overall	x→en	en→x	overall	x→en	en→x	overall	x→en	en→x
WΣER	0.616	**0.694**	0.541	55.1	50.0	61.2	39.1	48.6	36.2
BΣER	0.629	0.666	0.594	60.3	55.2	**68.8**	46.1	39.5	**52.5**
WBΣER	**0.639**	**0.696**	0.585	**68.0**	**67.1**	63.7	**48.7**	**52.6**	45.0
XENΣER	0.587	**0.692**	0.487	51.9	63.2	41.2	37.8	**52.6**	23.7
ENXΣER	0.599	0.595	**0.602**	50.6	38.1	62.5	39.1	32.9	45.0

Table 3: *mean*, *rank*$\geq$ and *rank*$>$ values for error rate sums compared over all documents, over English documents and over non-English documents.

target language separately, filtering error categories by POS classes, etc.

Acknowledgments

This work has partly been developed within the TARAXÜ project[4] financed by TSB Technologiestiftung Berlin – Zukunftsfonds Berlin, co-financed by the European Union – European fund for regional development. Special thanks to Mark Fishel and Ondřej Bojar.

References

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further Meta-Evaluation of Machine Translation. In *Proceedings of the 3rd ACL 08 Workshop on Statistical Machine Translation (WMT 2008)*, pages 70–106, Columbus, Ohio, June.

Chris Callison-Burch, 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 (WMT 2009)*, pages 1–28, Athens, Greece, March.

Chris Callison-Burch, 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 2010)*, pages 17–53, Uppsala, Sweden, July.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 Workshop on Statistical Machine Translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT 2011)*, pages 22–64, Edinburgh, Scotland, July.

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

Maja Popović and Hermann Ney. 2011. Towards Automatic Error Analysis of Machine Translation Output. *Computational Linguistics*, 37(4):657–688, December.

Maja Popović. 2011. Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics*, (96):59–68, October.

Drahomíra Spoustová, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. 2007. The best of two worlds: Cooperation of statistical and rule-based taggers for czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, ACL 2007*, pages 67–74, Praha.

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