Correlating decoding events with errors in Statistical Machine Translation

Eleftherios Avramidis **Maja Popović**
German Research Center for Artificial Intelligence (DFKI GmbH)
Language Technology Lab
Alt Moabit 91c, 10559 Berlin
eleftherios.avramidis@dfki.de maja.popovic@dfki.de

Abstract

This work investigates situations in the decoding process of Phrase-based SMT that cause particular errors on the output of the translation. A set of translations post-edited by professional translators is used to automatically identify errors based on edit distance. Binary classifiers predicting the sentence-level existence of an error are fitted with Logistic Regression, based on features from the decoding search graph. Models are fitted for 3 common error types and 6 language pairs. The statistically significant coefficients of the logistic function are used to analyze parts of the decoding process that are related to the particular errors.

1 Introduction

Evaluating the output of Machine Translation (MT) has been in the focus since the first developments of the field. There have been several efforts to measure the translation performance, or to identify errors by defining manual and automatic metrics.

Advanced automatic metrics and Quality Estimation methods have introduced machine learning (ML) techniques in order to predict indications about the quality of the produced translations (Lavie and Agarwal, 2007; Stanojevic and Sima'an, 2014). When compared to traditional automatic metrics, ML techniques allow acquiring knowledge about the quality of the translation out of a big amount of features. Such features are typically *black-box* features, generated by automatic analysis over the text of the source or the translations, or less often *glass-box* features, derived from the internal functioning of the translation mechanism.

In this work, we focus on the glass-box features. However, instead of focusing on the performance of a quality assessment mechanism, we look backwards into what happened during the decoding process and led into known errors in the translation output.

2 Problem definition

This work uses ML in order to fit a statistical model, associating properties and events of the decoding process with the existence of particular errors of a phrase-based statistical MT system. Such a model optimizes a function f:

$$Y = f(X) = \beta \circ X \tag{1}$$

where:

- X is the independent variable, i.e. a feature vector representing properties and events of the decoding process and has been extracted from the decoding search graph
- Y is the dependent variable, i.e a value predicting the existence of errors
- β is a weight vector estimated by ML to minimize the error of the function $\circ$, given samples of X and Y. This vector contains coefficients for each one of the features. Given a well-fit model and a relevant statistical function, these coefficients can indicate the importance of each feature.

Our aim is to use the β coefficients in order to explain several behaviours of the decoding process, relevant to the errors. The exact formulation of the statistical function $\circ$ is given in Section 4.3.

Our intention is to not train the model using as a dependent variable a complex quality metric such as BLEU (Papineni et al., 2002) or WER, since this would increase complexity by capturing many issues in just one number. Instead, we choose a more fine-grained approach, by focusing onto specific type of errors that occur often in machine translation output (Section 4.1).

3 Related Work

Error detection for MT and Quality Estimation is an important component of post-editing approaches. Our work focuses solely on features derived from the decoding process.

The first experiments on "Confidence Estimation" make use of a small number of Statistical Machine Translation (SMT) features in order to train a supervised model for predicting the quality of the Translation (Blatz et al., 2004). Later work, identified as "Quality Estimation", defines such features as "glass-box" features (Specia et al., 2009). 54 glass-box features are shown to be very informative, when fitted in a regression model, along with other black-box features.

Avramidis (2011) uses decoding features in a sentence-level pairwise classification approach for Hybrid MT in order to select the best translations out of outputs produced by statistical and rule-based systems, whereas a corpus of machine translation outputs with internal meta-data was released at that time (Avramidis et al., 2012). Later works use glass-box features in order to predict numerical indications of translation quality, such as post-editing effort (Rubino et al., 2013; Hildebrand and Vogel, 2013) or post-editing time (Avramidis et al., 2013). Contrary to these works, we only predict specific error types, with the focus on understanding the contribution of the features.

Prediction of specific error types was included in the shared tasks of the 8th and 9th Workshop on Statistical Machine Translation (Bojar et al., 2013; Bojar et al., 2014). Several participants contributed systems that predict error types (Besacier and Lecouteux, 2013; Bicici and Way, 2014; de Souza et al., 2014). In that case, prediction was done on the word level and contrary to our experiments, no glass-box features were used, therefore there was no connection of the ML with the decoding process.

Guzmán and Vogel (2012), in the work that is most related to ours, aim to identify the contribution of the features. Similar to several previously mentioned works, a multivariate linear regression model is trained in order to predict continuous quality values of complex metrics. Although the aim of this work is similar to ours, we work in a more fine-grained way: instead of modelling metrics, we try to explain the contribution of the decoding features on the occurrence of specific error types.

4 Methods

4.1 Error detection

The errors on the target output are used as a training material for the supervised ML algorithm. They are identified as a subset of commonly observed error categories (Vilar et al., 2006). For our purpose we focus on *missing words*, *extra words* and *re-ordering errors*, since our data give sufficient amounts for training a statistical model on these error categories.[1].

In order to detect the errors on the translation output, we follow the automatic method by Popović and Ney (2011), which has shown to correlate well with human error annotation. This method automatically detects errors based on the edit distance of the produced translation against a reference human translation. An example of how errors are detected can be seen in Figure 1.

4.2 Phrase-based SMT search graph

The glass-box features are extracted from the decoding process of a phrase-based SMT system (Koehn et al., 2003) with cube-pruning (Huang and Chiang, 2007). The decoding process performs a search in various dimensions, calculating scores for many phrases and hypothesis expansions. Most scores are difficult to be interpreted as glass-box features in their initial form. The amount of scores calculated per sentence is not fixed, whereas the basic requirement for each feature is to have only one value that is valid on a sentence level, so that it can be used in the sentence error prediction model.

For this purpose, we process the verbose output of the decoder and derive scores, counts and other statistics that can have this sentence-level interpretation. When decoding steps contain a number of scores which is not fixed for every sentence, we extract features out of their statistics, such as the mean and the standard deviation, the minimum/maximum value and their position in the sentence. An example of how some of these features are extracted is illustrated in Table 1. On the upper part of the table, one can see the log-probability and the future cost estimate for each one of the phrases in the sentences. On the lower part we demonstrate some statistics that are derived from the scores and the positions of the words in the upper part.

[1] Although previous work defines 5 error types, not all of them could be sufficiently modelled given this amount of data

source:	Überraschenderweise zeigte sich, dass die neuen Räte in Bezug auf diese neuen Begriffe etwas im Dunkeln tappen.
translation:	Surprisingly, showed that the new councils in relation to these new concepts slightly in the dark.
post-editing:	Surprisingly, [miss:it] [lex:seems] that the new [lex:councillors] [miss:are] [reorder: slightly in the dark] in relation to these new concepts.

Figure 1: Example of the results found with the automatic error detection process. One can see missing words, reordering of 4 words and some lexical errors, which are not discussed in this work

Similar practice is applied to extract the entire set of 104 glass-box features, which includes:

Phrase counts and positions: The produced translation consists of sets of phrases that are chosen as the most probable hypothesis. On this hypothesis we count the number of phrases, words, the length of the phrases, the length difference between source and aligned target phrases and also the position of the shortest and the longest phrase in the sentence.

Unknown tokens are words or phrases that are not found in the phrase table. Their count, their ratio and their position in the translated sentence (average position, standard deviation of their positions, position of first and last unknown word) are included as features.

Translation probabilities: *Log probability* (pC) and *future cost estimate* (c) are available for each phrase of the chosen hypothesis. We extract their average, standard deviation and also their minimum and maximum values and their position in the sentence. Additionally, we count phrases whose pC or c is too low or too high. This is done by checking whether their values are out of the standard deviation of all phrases in the sentence.

Time: The decoder reports the time required for the entire translation process, the search, the language model calculation, the generation of hypotheses other than the ones chosen and for collecting translation options. We use these as features, also averaged over the entire translation time.

Decoding graph: These features come from the entire set of alternative phrase hypotheses generated during the search. From the entire set of alternative hypotheses he derive statistics for their log probability, the future estimate (average, standard deviation, count of alternative phrase hypotheses lower and higher than the standard deviation).

4.3 Machine Learning

The existence of an error (binary classification) is modelled with **logistic regression**. It is a widely-used ML method that optimizes a logistic function to predict values Y in the range between zero and one (Cameron, 1998), given a feature set X:

$$Y = \beta \circ X = \frac{1}{1 + e^{-1(a+\beta X)}} \qquad (2)$$

For fitting the model we use the Newton-Raphson algorithm, which minimises iteratively the least-squares error given the training data (Miller, 2002). The regression fitting included *Stepwise Feature Set Selection* (Hosmer, 1989).

In order to assess the contribution of individual predictors in a given model, we examine the significance by calculating a p-value for each of them. This is the probability that the beta coefficient differs from 0.0. The probability is computed based on Wald statistic of each co-efficient, following the χ^2 distribution. The Wald statistic is the ratio of the square of the regression coefficient to the square of the standard error of the coefficient (Menard, 2002; Harrell, 2001).

5 Experiment

5.1 Data and models

For the purpose of our analysis we train one logistic regression model per error category and language pair, practically resulting into 21 models. We include a set of models trained on the data from all the language pairs, in order to model cases that are independent of the languages involved in the translation, or that are not statistically significant in single language pairs, due to data sparsity.

The experiment is based on data from WMT11 (Callison-Burch et al., 2011), augmented with a small amount of data of WMT10 (Callison-Burch et al., 2010) and technical documentation of mechanical engineering equipment, as provided

source: [überraschenderweise] [zeigte sich] [, dass die neuen] [Räte] [in Bezug auf] [diese neuen] [Begriffe] [etwas] [im Dunkeln tappen .]

translation: [surprisingly ,] [showed] [that the new] [councils] [in relation to] [these new] [concepts] [slightly] [in the dark .]

position	phrase	pC	c
[0..0]	surprisingly ,	−0·770335	−2·69341
[1..2]	showed	−1·54184	−2·81277
[3..6]	that the new	−0·563381	−2·65923
[7..7]	councils	−0·386571	−1·98291
[8..11]	in relation to	−1·29663	−2·85591
[12..13]	these new	−0·332607	−2·17422
[14..14]	concepts	−0·540415	−2·01213
[15..15]	slightly	−0·585549	−2·00382
[16..19]	in the dark .	−1·48327	−3·90992
minimum		−1·54184	−3·90992
maximum		−0·332607	−1·98291
average (avg)		−0·83334	−2·56715
standard deviation (std)		0·448	0·5862
no of phrases with score lower than avg-std		3	1
no of phrases with score higher than avg+std		1	0
averaged position of phrase with lowest score		0·11111	0·88889
averaged position of phrase with highest score		0·55556	0·33333

Table 1: Example glass-box feature extraction from the decoding result. Decoding scores such as phrase *log probability* (pC) and *future cost estimate* (c), whose number is not the same for every sentence (upper part of the table), are reduced to a fixed feature vector based on basic statistics (shown on the lower part of the table)

by a translation agency. The amount of sentences originating from each data source per language pair is shown in Table 2. Almost half of these sentences are given to professional translators, with the instructions to perform as few changes as possible in order to correct the translations. The size of the corpus and the number of post-edited (p.e) sentences can be seen in Table 3.

Minimal post-editing is considered to be ideal for automatic error detection. In contrast, reference translations may contain severe alterations to the structure of the sentence, misleading the automatic error detection. Nevertheless, as it can be seen in 3, the amount of sentences for certain error types and language pairs may be too small, leading to a severely sparse set of training data and therefore weak models. Consequently, as it was not technically possible to acquire post-editing on a bigger amount of data, we perform error-detection on a mixture of post-editions and reference translations, in case this increases the quality of the statistical models. Preliminary experiments confirmed the positive effect, as the precision and recall was increased on most models (even up to 29%) when adding errors detected on reference translations to the ones detected on post-editing. Despite some obvious drawbacks, this move is also motivated by the fact that the original experiments that showed the accuracy of the automatic error detection (Popović, 2011a) were also performed against reference translations.

5.2 Experiment set-up

As a statistical phrase-based system we trained one Moses (Koehn et al., 2007) system per language direction, using Europarl (Koehn, 2005) and News Commentary corpora[2]. Its settings follow the WMT11 baseline, including a compound splitter for German-English and a truecaser for all language pairs. The system was tuned

[2] News Commentary was used only for German-English due to lack of alignments for the other languages

	de-en	de-fr	de-es	en-de	fr-de	es-de
wmt10	118	30	33	14	74	0
wmt11	952	952	80	1087	977	101
customer	741	0	430	0	0	830
total	1811	982	543	1101	1051	931

Table 2: Amount of senteces from various sources per language pair

lang	sentences		reordering err.		missing words		extra words	
	total	p.e	total	p.e	total	p.e	total	p.e
de-en	1811	1139	1043	474	1079	570	869	454
en-de	1101	315	891	232	671	151	722	208
de-fr	982	198	819	157	597	80	630	147
fr-de	1051	122	851	88	691	76	621	66
de-es	543	543	288	288	322	322	186	186
es-de	931	931	345	345	333	333	339	339
all	6419	3248	4237	1584	3693	1532	3367	1400

Table 3: The size of the corpus per error category and language pair. p.e. indicates the number of sentences that were minimally post-edited by professional translators

with MERT using the news corpus test set from WMT07 (Callison-Burch et al., 2007). The decoding features are extracted from Moses' verbose output of level 2. Our target language model with an order of 5 is trained with SRILM toolkit (Stolcke, 2002), based on the respective monolingual training material. The Orange toolkit (Demšar et al., 2004) is used for processing and running the Logistic Regression algorithms. The Hjerson tool (Popović, 2011b) was used in order to detect errors on the translation.

6 Results

6.1 Model performance

A necessary step is to check how well each model fits the data, since a well-fit model is required for drawing conclusions. For this purpose we perform cross-fold validation with 10 folds. The precision and recall scores are shown in Table 4. Precision indicates the ratio of the predicted sentences that contain an error, whereas recall indicates the ratio of the sentences that have an error and are successfully predicted.

The model predicting the existence of reordering errors has the highest precision and recall on all individual language pairs and achieves a generally high precision of about 83-87% (apart from Spanish-German). The model of predicting missing words seems most successful on the dataset combining all language pairs. Extraneous words have models with much lower scores, which means that it is more difficult to draw conclusions.

6.2 Analysing coefficients

We proceed to the analysis by considering the β coefficient of the fitted logistic function (function 2) for each feature. Additionally, the confidence p-value indicates the evidence that the respective feature contributes to the prediction of the outcome.

The feature coefficients given by the fitted logistic function vary per error category and language pair. This is understandable, given the fact that the translation systems and the test sentences are quite heterogeneous among language pairs.

Interpretation of the feature coefficients may vary. The most clear indication is the positive or negative sign of each coefficient. Additionally, one has to note that several features result as a mathematical function of other features; thus, when they all occur in the logistic function, explaining the coefficients of a feature should not neglect the existence of the features that are mathematically related to it.

In order to lead to useful conclusions, we show the feature coefficients who seem to have a statistically significant relation to known functionality of the decoding process for several language pairs. Conclusions are detailed per error category, based on Tables 5, 6 and 7 which include the beta co-

	all		de-en		de-fr		de-es		en-de		fr-de		es-de	
	prec	rec	prec	rec	prec	rec	prec	rec	prec	rec	prec	rec	prec	rec
ext	.70	.70	.68	.59	.72	.82	.61	.52	.73	.85	.67	.76	.62	.53
miss	.85	.88	.75	.76	.67	.79	.80	.82	.67	.78	.70	.86	.64	.52
reord	.70	.79	.83	.81	.88	.96	.85	.82	.85	.93	.87	.92	.77	.70

Table 4: Precision (prec) and recall (rec) of the logistic regression model, measure with cross-validation. There is one model per combination of language pair and error category; plus one model trained on data from all language pairs per error category. High precision and recall indicate that the model is well-fit.

efficient and the respective p-value for statistical significance. Coefficients not appearing in specific cells of the table have been eliminated by the stepwise feature set selection. For a few language pairs we also include coefficients with p-values higher than our confidence level, when the same coefficient is statistically significant for some other language pairs.

6.2.1 Reordering errors

The coefficients for the reordering errors are shown in Table 5. First, the sentence-level ratio of unknown words (unk-per-tokens) the standard deviation of the **position of an unknown token** in the translated sentence (unk-pos-std) has a positive effect towards the creation of a reordering error. A high standard deviation means that unknown tokens are scattered in distant places along the sentence; being "unknown" they cannot be captured by the lexical reordering model and it is therefore likely that they cause an erroneous phrase order in the parts of the sentence where they occur. This is confirmed for 3 of our models (Table 5) with p-value p <0.10, including the model trained on all language pairs.

Calculating the **length difference between the source phrases and the aligned target phrases** chosen for the translation output provides a useful feature too. The maximum difference of all source-phrase lengths minus the respective chosen target-phrase lengths (src-tgt-diff-max) has a positive effect into creating a reordering error. This effectively occurs for cases when the decoder chooses to translate a source phrase with a much shorter target phrase. Invertedly, the smallest difference between source and target phrase (src-tgt-diff-min) has a negative effect on creating reordering errors. The two indications are confirmed with a positive statistically significant coefficient for 6 cases in our models.

All of our translation systems into German increase their reordering errors due to a big **length of the longest source phrase** chosen by the decoder (src-phrase-len-max). This may be due to the fact that German language has significantly different word order than the other languages and this results into a common error for SMT systems.

6.2.2 Missing words

For the language pairs indicated in Table 6, there is a positive effect on having missing words by the **standard deviation of the phrase log probability** (pC-std) and the **position of the phrase with the lowest log probability** (pC-min-pos) towards the end of the sentence, confirmed by our logistic regression models for 4 language pairs. Similar are the conclusions for the **future cost estimate** (c-avg) averaged to the number of phrases, and the **number of phrases with low future cost estimate**, i.e. when it is lower than the mean of all phrases minus the standard deviation. These features reflect situations during the decoding when dropping a phrase results to a higher overall probability. This may occur due to the low probability and the high future cost estimate of all the possible translations of the phrase.

In other cases, there may be words missing when the phrase alignments chosen for translating the source sentences are longer, namely when **the average source length** (src-phrase-len-avg) is higher and the **length of the shortest source phrase** (src-phrase-len-min) is lower. A positive impact to having words missing is also given by the **standard deviation of the length difference between respective source and target phrases** (src-tgt-diff-std), i.e. when the length between aligned source and target phrases varies a lot.

6.2.3 Extra words

The coefficients from the models on extra words are not so conclusive, due to their low precision and recall. One can note that the **time for calculating the language model** and the **total time**

	all		de-en		de-fr		de-es		en-de		fr-de		es-de	
	β	p	β	p	β	p	β	p	β	p	β	p	β	p
unk-per-tokens	2.08	.00									5.98	.07	3.03	.19
unk-pos-std	0.19	.00	0.22	.00					0.17	.15	0.25	.09		
pC-var	-1.42	.06	-0.70	.37	-10.66	.00					-9.95	.01		
src-phrase-len-max									-0.50	.04	-0.36	.17	-1.45	.02
src-tgt-diff-min	-0.26	.02					-0.60	.00	-0.50	.06	-0.84	.01		
src-tgt-diff-max	0.27	.02	0.14	.30	0.24	.21			0.83	.01	0.47	.13		

Table 5: Indicative beta coefficients and their respective p-values for the features affecting the existence of reordering errors.

	all		de-en		de-fr		de-es		en-de		fr-de		es-de	
	β	p	β	p	β	p	β	p	β	p	β	p	β	p
pC-min-pos	0.40	.00	0.57	.02			1.06	.03					1.15	.01
pC-std	2.11	.00	1.14	.50	3.82	.13	9.46	.00					7.05	.02
c-avg	0.52	.02	0.66	.08					1.44	.17	2.66	.07	2.14	.00
c-low			.11	.20	.19	.04	0.28	.18						
src-phrase-len-avg	0.70	.00	1.23	.00	1.56	.01							1.74	.01
src-phrase-len-min	-0.48	.08	-0.88	.13	-0.86	.30					-1.10	.09	-0.95	.38
src-tgt-diff-std	0.41	.02	0.87	.04	0.91	.08					1.06	.02	0.57	.16

Table 6: Indicative beta coefficients and their respective p-values for the features related to the error of missing words.

of the translation (time-translation) get increased when extra words occur, for some language pairs. The effect of the **standard deviation of the unknown tokens** (unk-pos-std) is opposite to what it causes to the reordering errors: the closer the unknown tokens are to each other, the more extra words occur.

For some models, including the model built on all language pairs, extra words correlate with the **length of the source sentence** (total-source-words), particularly when translating from English and French into German. Three models also indicate small but highly significant contribution by a feature from the search graph: the count of **alternative hypothesized phrases, whose log probability is lower** than the standard deviation of the log probability in their respective sentence (alt-pC-low).

7 Conclusions and further work

We have provided statistical evidence for how the functioning of the phrase-based SMT decoding process affects the existence of three frequently occurring error types. The existence of the errors in a sentence is modelled over some decoding process features with logistic regression, which resulted into several models with satisfactory precision and recall values.

By grouping the observations by error type, we managed to show important features (representing stages of the decoding process) that are common for several language pairs at the same time. Most of the indications observed are based on statistically significant coefficients.

One observation is that the chosen method is traditionally employed to examine feature contributions in a specific model, which is seldom generalized across different models. Moreover, although the features in the decoding procedure do affect the translation performance, there are concerns that the logistic relationship between decoding features and specific translation errors is very large, so that the statistical relationship is hard to be captured by simple binary classification approaches. Our next efforts will therefore look on other machine learning methods, also considering the possibility to model the amount and/or the exact location of errors.

Further work could extend this effort by including a wider range of error categories that describe better the requirements for a translation correct output. Instead of automatically detecting errors on post-edited output, a possible extension could consider modelling error types assigned by humans. Additionally, the analysis of features can be extended in order to cover other types of machine translation, such as hierarchical phrase-based translation and rule-based translation.

An obvious application of this analysis would be incorporating the findings into the decoding

	all		de-en		de-fr		de-es		en-de		fr-de		es-de	
	β	p	β	p	β	p	β	p	β	p	β	p	β	p
pC-std	1.55	.09	2.06	.17									7.03	.03
pC-max-pos	0.38	.00	0.46	.04			1.29	.01						
time-calculate-lm	0.89	.02												
time-translation	0.52	.34			2.35	.06	0.27	.37	2.07	.10				
unk-pos-std	-0.02	.16					-0.11	.11			-0.21	.01		
total-source-words	0.22	.01							0.18	.00	0.27	.01		
alt-pC-low	0.01	.03	0.02	.06	0.04	.00								

Table 7: Indicative beta coefficients and their respective p-values for the features affecting the existence of erroneous extra words.

process, in order to improve it, e.g. by introducing features to the decoding engine that directly indicates the factors that cause errors.

Acknowledgment

This work has received support by the EC's FP7 (FP7/2007-2013) under grant agreement number 610516: "QTLeap: Quality Translation by Deep Language Engineering Approaches".

References

Eleftherios Avramidis, Marta R Costa-Jussà, Christian Federmann, Josef van Genabith, Maite Melero, and Pavel Pecina. 2012. A Richly Annotated, Multilingual Parallel Corpus for Hybrid Machine Translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2189–2193, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Eleftherios Avramidis, Maja Popović, and Maja Popovic. 2013. Machine learning methods for comparative and time-oriented Quality Estimation of Machine Translation output. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 329–336, Sofia, Bulgaria, August. Association for Computational Linguistics.

Eleftherios Avramidis. 2011. DFKI System Combination with Sentence Ranking at ML4HMT-2011. In *Proceedings of the International Workshop on Using Linguistic Information for Hybrid Machine Translation and of the Shared Task on Applying Machine Learning Techniques to Optimising the Division of Labour in Hybrid Machine Translation*, pages 99–103, Barcelona, Spain. Center for Language and Speech Technologies and Applications (TALP), Technical University of Catalonia.

Laurent Besacier and Benjamin Lecouteux. 2013. LIG System for WMT13 QE Task : Investigating the Usefulness of Features in Word Confidence Estimation for MT. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*.

Ergun Bicici and Andy Way. 2014. Referential Translation Machines for Predicting Translation Quality. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 313–321, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *8th Workshop on Statistical Machine Translation*, Sofia, Bulgaria, August. Association for Computational Linguistics.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

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

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 Metrics*, pages 17–53, Uppsala, Sweden, July. Association for Computational Linguistics.

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*, pages 22–64, Edinburgh, Scotland, July. Association for Computational Linguistics.

Adrian Cameron. 1998. *Regression analysis of count data*. Cambridge University Press, Cambridge UK; New York NY USA.

José Guilherme de Souza, Jesús González-Rubio, Christian Buck, Marco Turchi, and Matteo Negri. 2014. FBK-UPV-UEdin participation in the WMT14 Quality Estimation shared-task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 322–328, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Janez Demšar, Blaž Zupan, Gregor Leban, and Tomaz Curk. 2004. Orange: From Experimental Machine Learning to Interactive Data Mining. In *Principles of Data Mining and Knowledge Discovery*, pages 537–539.

Francisco Guzmán and Stephan Vogel. 2012. Understanding the Performance of Statistical MT Systems: A Linear Regression Framework. In *Proceedings of COLING 2012*, pages 1029–1044, Mumbai, India, December. The COLING 2012 Organizing Committee.

Frank E Harrell. 2001. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer.

Silja Hildebrand and Stephan Vogel. 2013. MT Quality Estimation: The CMU System for WMT'13. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 373–379, Sofia, Bulgaria, August. Association for Computational Linguistics.

David Hosmer. 1989. *Applied logistic regression*. Wiley, New York [u.a.], 8th edition.

Liang Huang and David Chiang. 2007. Forest rescoring: Faster decoding with integrated language models. In *Annual Meeting-Association For Computational Linguistics*, volume 45, page 144.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 177–180, Prague, Czech Republic, June.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. *Proceedings of the tenth Machine Translation Summit*, 5:79–86.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic, June. Association for Computational Linguistics.

Scott Menard. 2002. *Applied logistic regression analysis*, volume 106. Sage.

Alan Miller. 2002. *Subset Selection in Regression*. Chapman & Hall, London, 2nd edition.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

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

Maja Popović. 2011a. From human to automatic error classification for machine translation output. In *15th International Conference of the European Association for Machine Translation*, Leuven, Belgium. European Association for Machine Translation.

Maja Popović. 2011b. Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics*, 96(-1):59–68.

Raphaël Rubino, Antonio Toral, Santiago Cortés Va\'illo, Jun Xie, Xiaofeng Wu, Stephen Doherty, Qun Liu, and Santiago Cortés Vaíllo. 2013. The CNGL-DCU-Prompsit Translation Systems for WMT13. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 213–218, Sofia, Bulgaria, August. Association for Computational Linguistics.

Lucia Specia, M. Turchi, N. Cancedda, M. Dymetman, and N. Cristianini. 2009. Estimating the Sentence-Level Quality of Machine Translation Systems. In *13th Annual Meeting of the European Association for Machine Translation*, pages 28–35, Barcelona, Spain.

Milos Stanojevic and Khalil Sima'an. 2014. BEER: BEtter Evaluation as Ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 414–419, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Andreas Stolcke. 2002. SRILM – An Extensible Language Modeling Toolkit. In *Proceedings of the Seventh International Conference on Spoken Language Processing*, pages 901–904. ISCA, September.

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