Automatic Post-Editing based on SMT and its selective application by Sentence-Level Automatic Quality Evaluation

Hirokazu Suzuki
Knowledge Media Laboratory
Corporate Research & Development Center
Toshiba Corporation, Japan
hirokaz.suzuki@toshiba.co.jp

Abstract

In the computing assisted translation process with machine translation (MT), post-editing costs time and efforts on the part of human. To solve this problem, some have attempted to automate post editing. Post-editing isn't always necessary, however, when MT outputs are of adequate quality for human. This means that we need to be able to estimate the translation quality of each translated sentence to determine whether post-editing should be performed. While conventional automatic metrics such as BLEU, NIST and METEOR, require the golden standards (references), for wider applications we need to establish methods that can estimate the quality of translations without references. This paper presents a sentence-level automatic quality evaluator, composed of an SMT phrase-based automatic post-editing (APE) module and a confidence estimator characterized by PLS (Partial Least Squares) regression analysis. It is known that this model is a better model for predicting output variable than a normal multiple regression analysis when the multicollinearity exists between the input variables. Experiments with Japanese to English patent translations show the validity of the proposed methods.

1 Introduction

The translation quality of MT has been improving but has not reached the adequate level compared with human translation. As such, manual evaluation and post-editing constitute an essential part of the translation processes. To make the best use of MT, human translators are urged to perform post-editing efficiently and effectively. Therefore there is a huge demand for MT to alleviate the burden of manual post-editing.

Since Rule-based MT (RBMT) is generally more stable in translation quality than Statistical MT (SMT), it can make it easier to integrate the post-editing into the translation processes. This, however, is also a weakpoint of RBMT because post-editors are forced to repeatedly correct the same kind of errors made by MT systems.

Recognizing that SMT is better suited to correct frequent errors to appropriate expressions, some (Simard, Goutte, & Isabelle, 2007) (Lagarda, Alabau, Casacuberta, Silva, & Diaz-de-Liano, 2009) have proposed to use SMT for an automatic post-editor and built an automatic post-editing module, where MT outputs are regarded as source sentences and manually post-edited/translated results as target sentences.

The rest of this paper is organized as follows. Section 2 shows an automatic post-editing (APE) module based on phrase-based SMT (Koehn, Och, & Marcu, 2003), which was trained with data from the Japanese-to-English patent translation task in

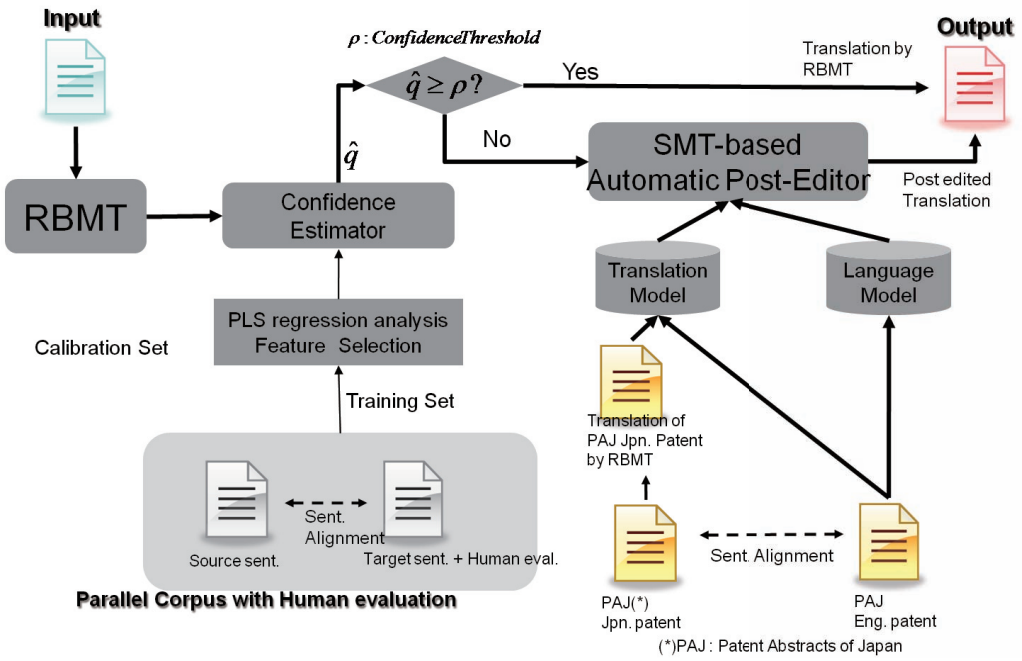

Fig. 1: SMT-based Automatic Post-editing system with confidence estimator

NTCIR7 (NII:The 7th NTCIR Workshop, 2007/2008) .

Section3 describes how to estimate translation quality at the sentence level without references by means of PLS regression analysis, found to be effective for Confidence Estimation (Specia, Cancedda, Turchi, & Cristianini, 2009) (Specia, Saunders, Turchi, Wang, & Shawe-Taylor, 2009).

Finally Section 4 shows the validity of the proposed method by examining the transitional changes in NIST scores when the APE is applied selectively according to the result of the estimator.

2 Automatic Post-Editing based on SMT

Fig. 1 shows the overview of the proposed system, which evaluates Japanese-to-English RBMT quality and performs automatically post-editing when the estimated quality is lower than some threshold.

The description in this section corresponds to 'SMT-based Automatic Post-Editor' in Fig. 1. 'Confidence Estimator' in Fig. 1 is described in Section 3.

I used Moses (Koehn, et al., 2007) as a phrase-based SMT module. For data setting, I extracted from about 1.80M sentence pairs in the Japanese-to-English parallel corpus (original corpus) as provided by the NTCIR7 patent translation task, about 1.18M sentence pairs excluding long sentences. Then I translated the extracted 1.18M Japanese sentences using RBMT to English and trained the translation model of SMT with the parallel corpus consisting of those RBMT results (1.18M RBMT English sentences) and the corresponding original English sentences. The language model was trained with 1.80M English sentences from the original corpus.

Here is the breakdown:

Model		# of Sentences	# of words	
			RBMT	Reference
Translation	training	1,184,827	33,719,825	33,356,416
	dev	805	26,277	25,681
Language		1,798,571		59,429,838

Table 1: Statistics of the training/dev corpus

The parameter settings at the training stage were as follows:

- Language model : 3-gram in SRILM, -order 3 -interpolate -kndiscount options were specified in 'ngram-count'.
- Translation model : -alignment grow-diag-final-and -reordering msd-bidirectional-fe options were specified.

For test data, I used 899 Japanese sentences from the test data in the NTCIR7 patent translation task.

Fig. 2 gives the NIST scores of RBMT results (raw translations) and SMT-based APE results,

where the latter scored about 15% higher than the former.

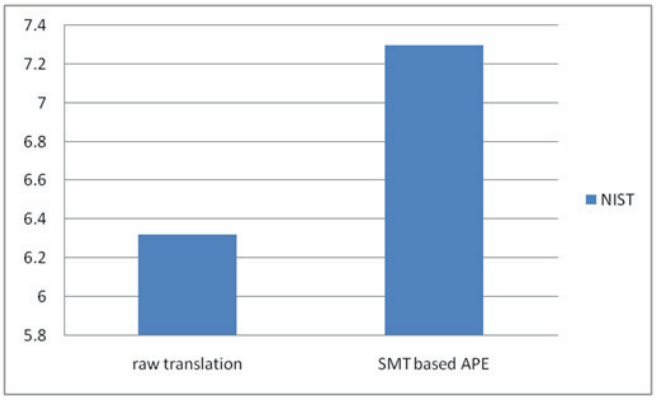

Fig. 2: NIST scores of raw translation and APE result

3 Automatic Quality Evaluation without References

The APE based on SMT can also be regarded as a domain adaptation for MT result, because the post-editing depends on the domain of the language model and the translation model. From this point of view, a selective application of the APE is preferred in order to avoid over-application to sentences that don't need to post-edit. In this case, we need the sentence-level evaluator of translation sentences.

The conventional automatic metrics, such as BLEU, NIST and METEOR, use similarity calculation and need references to calculate the similarities between system outputs and references. However, it is costly to prepare several references for each source sentence in a test data. Moreover, it is practically impossible to give all variations of translations.

Taking these problems into account, some studies use confidence estimation (CE) metrics to estimate translation quality at the sentence level without references. (Blatz, et al., 2003) consider CE as a binary classification of translation as either 'good' or 'bad', while (Specia, Cancedda, Turchi, & Cristianini, 2009) and (Specia, Saunders, Turchi, Wang, & Shawe-Taylor, 2009) consider CE as estimating a continuous translational quality score for each sentence, using PLS regression analysis.

In this paper, I show the feature set for evaluation of RBMT and the quality prediction model by PLS regression analysis with them for APE.

3.1 PLS Regression Analysis

First, I will briefly explain the main features of PLS regression analysis.

A regression analysis focuses on the relationship between a dependent variable (response variable) and one or more independent variables (explanatory variables) and estimates the behavior of the response variable given the explanatory variables. A regression analysis using a linear model with one explanatory variable is called 'simple linear regression' and one with two or more explanatory variables is called 'multiple linear regression (MLR)'. (Zhu, Yang, Wang, Wang, & Li, 2009) have constructed an automatic evaluator for human translations by means of MLR analysis, in which the response variable is a translation score given by human and the explanatory variables are numbers of errors at word/phrase/syntax/discourse levels.

However, MLR implicitly assumes that the explanatory variables are linearly independent, i.e. it is not possible to express any explanatory variable as a linear combination of the others. Therefore, if there is correlation between the explanatory variables, the estimation with MLR tends to fail eventually. This phenomenon is called 'multicollinearity'.

It is known that Partial Least Squares (PLS) regression analysis (Wold, Ruhe, Wold, & Dunn, 1984) is a better way to build an accurate prediction model even if multicollinearity is present as in the case of automatic evaluation of translation quality without references.

In evaluating translation quality automatically without references by regression analysis, we cannot assume that the linguistic features used as the explanatory variables are linearly independent, nor can we determine the linguistic features so that they are linearly independent.

In these situations, PLS regression analysis is likely able to build an accurate quality prediction model.

3.2 Building a Quality Prediction Model with PLS Regression Analysis

I have built quality prediction model with PLS regression analysis. First of all, we need to determine what the input variables are and what the output (response) variables are.

3.2.1 Input Variables

I defined the features to be used as input variables in PLS regression analysis as Table 2.

The median (second quartile, Q2) is the middle value of the data set $\{x_i\}$, which is arranged in ascending order of magnitude:

$$Q2 = x_{(n+1)/2}$$

, where n denotes the number of the data.
The lower quartile (first quartile, Q1) is the median of the lower half of the data set:

$$Q1 = x_{(n+1)/4}$$

The upper quartile (third quartile, Q3) is the median of the upper half of the data set:

$$Q3 = x_{3(n+1)/4}$$

And interquartile range (IQR) is used as a robust measure of statistical dispersion and is defined as:

$$IQR = Q3 - Q1$$

In this paper, the data set consists of n-gram frequencies in translation sentences from the monolingual corpus which was used to train the language model for SMT-based automatic post-editing module.

I used the TreeTagger (Schmid, 1994) as an English morphological analyzer and the Link Grammar Parser (Grinberg, Lafferty, & Sleator, 1999) as an English parser.

3.2.2 Response variables

The translation results from each participant in the NTCIR7 Japanese-English patent translation task had been already evaluated by 3 human raters (A, B and C) with 5-grade Adequacy/Fluency scores. These scores are suitable for the purpose to build quality prediction model because our final goal is to alleviate the burden of human post-editor.

Therefore, I used these scores as the response variables in PLS regression analysis, i.e. eventually adequacy prediction model and fluency prediction model have been built.

3.2.3 Inconsistency between Human Raters

It is known that Adequacy/Fluency evaluation tends to be inconsistent among raters. The human evaluations in the NTCIR7 Japanese-English patent translation task is no exception.

Feature ID	Feature Description
1	(IQR of 1gram)/(total # of 1grams)
2	(Q3 of 1gram - Q2 of 1gram)/(Q2 of 1gram - Q1 of 1gram)
3	(Q2 of 1gram)/(total # of 1grams)
4	(IQR of 2gram)/(total # of 2grams)
5	(Q3 of 2gram - Q2 of 2gram)/(Q2 of 2gram - Q1 of 2gram)
6	(Q2 of 2gram)/(total # of 2grams)
7	(IQR of 3gram)/(total # of 3grams)
8	(Q3 of 3gram - Q2 of 3gram)/(Q2 of 3gram - Q1 of 3gram)
9	(Q2 of 3gram)/(total # of 3gram)
10	(Sum of MI(mutual information) between nonadjacent translational words)/(total # of word pairs)
11	(Sum of Dice Coefficient between nonadjacent translational words)/(total # of word pairs)
12	2-gram language model probability
13	3-gram language model probability
14	2-gram backward language model probability
15	3-gram backward language model probability
16	POS 2-gram language model probability
17	POS 3-gram language model probability
18	POS 2-gram backward language model probability
19	POS 3-gram backward language model probability
20	(# of nouns)/(# of words)
21	(# of verbs)/(# of words)
22	(# of adjectives)/(# of words)
23	(# of adverbs)/(# of words)
24	# of valid linkages by Link Grammar Parser
25	# of invalid linkages by Link Grammar Parser
26	(# of null counts by Link Grammar Parser)/(# of word)
27	# of noun phrases by Link Grammar Parser
28	# of verb phrases by Link Grammar Parser

Table 2: Feature Set

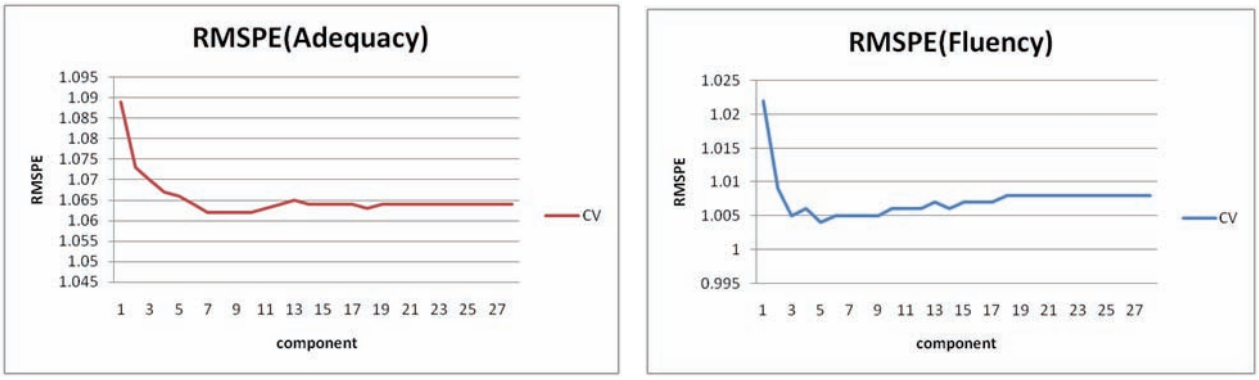

Fig. 3: The number of the latent variables and RMSPE in Adequacy/Fluency prediction

Fig. 5 shows Adequacy/Fluency evaluations by 3 human raters for some system. As for adequacy, the scores are similar to one another.

But rater A's fluency evaluations are very different from other two raters', in some sentences, almost opposite.

Therefore I can't use all of 3 raters' evaluation. In this paper, I don't use rater A's evaluations.

3.2.4 Training Set/Test Set

At the training stage, I selected at random 8 systems from all systems which took part in the NTCIR7 Japanese-English patent translation task. Then for each system, I got the features in Table 2 and the given human evaluations (5-grade Adequacy/Fluency score).

At the test stage, I selected at random 2 systems other than the above 8 systems and got their features and human evaluation results as well.

3.2.5 Latent Variables

PLS regression analysis does not perform regression analysis directly using input variables as the explanatory variables. It extracts the latent variables which affect the underlying relationship between input and output, and uses the latent variables as the explanatory variables.

In this paper, I computed the average error by means of Root Mean Squared Prediction Error (RMSPE):

$$\mathrm{RMSPE} = \sqrt{\frac{1}{n}\sum_{j=1}^{N}(y_j - \acute{y}_j)^2}$$

, where y_j denotes the observed value and $\acute{y}_j$ denotes the predicted value, and determined the number of the latent variables (components) by cross-validation (CV).

Fig. 3 shows the result of CV. Maximum performance in adequacy prediction is obtained with 8 components. In fluency prediction, with 5 components. At these number of components, RMSPE is minimum.

3.2.6 Prediction result

Table 3 shows the results by the adequacy prediction model and the fluency prediction model.

RMSPE of the fluency prediction is smaller than that of the adequacy prediction in both test systems. Speaman's rank correlation coefficients are as well. This is probably because the feature set defined as Table 2 depends only on the translational information as it is difficult to fairly judge how accurately the translations reflect the information in the source sentences.

Adequacy Prediction	Test System1	Test System2
RMSPE	1.07	1.48
Spearman's rank correlation coeff.	0.25	0.41
Fluency Prediction		
RMSPE	0.86	1.28
Spearman's rank correlation coeff.	0.30	0.37

Table 3 : Adequacy/Fluency prediction results

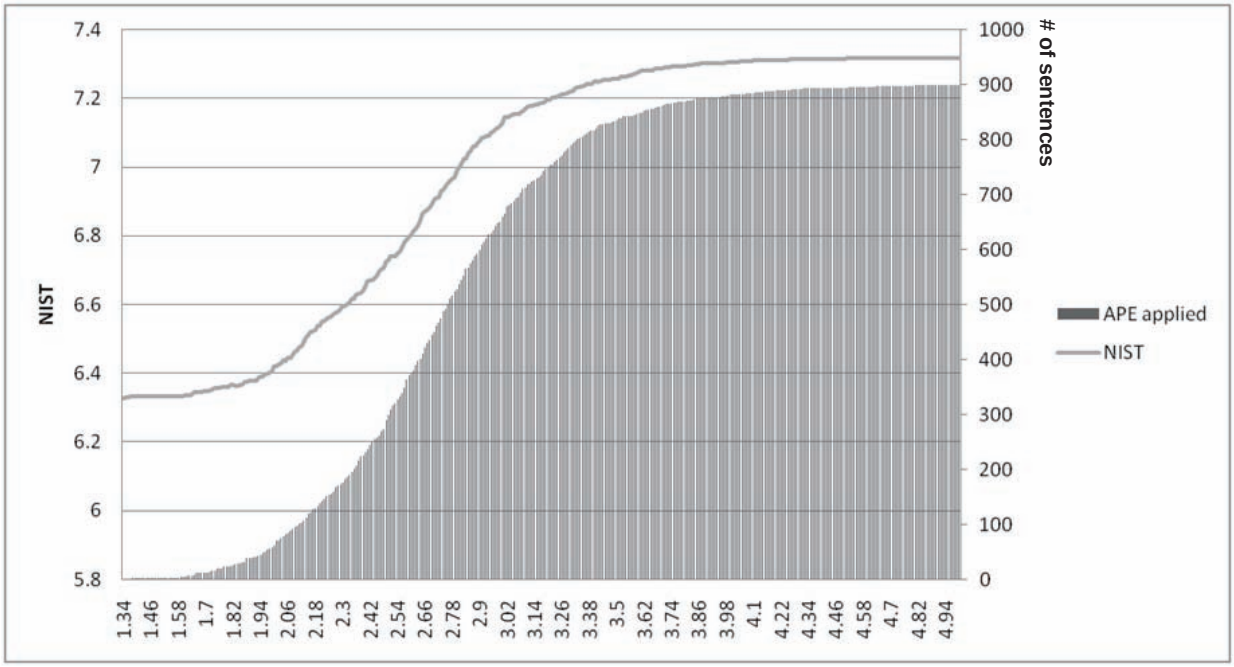

Fig. 4: Performance (NIST) and # of APE applied sentences for different Confidence thresholds

4 Integration of Quality Prediction Model into SMT-based APE

The proposed system automatically evaluates RBMT results by the quality prediction model based on PLS regression analysis, and decides whether SMT-based APE should be performed according to the evaluations obtained.

Predicted quality (confidence, $\hat{q}$) is calculated from the harmonic average between the adequacy prediction and the fluency prediction. As Fig. 1 shows, if the confidence is larger than threshold (confidence threshold) ρ, then the system outputs RBMT result with no change. Otherwise, the system post-edits the RBMT result, reckoning that it needs domain adaptation, then outputs the post-edited sentence.

4.1 Experiment

I tested the system on 899 Japanese sentences from the NTCIR7 J-to-E patent translation task as inputs.

Fig. 4 shows the transitional change of NIST (solid line) and the number of automatically post-edited sentences for the confidence threshold ranging from 1.34 (minimal value of $\hat{q}$) to 5.21 (maximum value of $\hat{q}$).

We can see both of transitional changes show a sharp rise for the threshold ρ between 2 and 3.5, but, on the other hand, the rise slows for the threshold ρ over 4.

If the confidence is higher than 4, it is highly probable that automatic post-editing is not necessary to perform.

5 Discussion and Future works

Applying the APE selectively according to the confidence reduces the unnecessary post-editing as illustrated by Table 4. Here, SRC denotes source sentence, REF a reference, RAW an RBMT result (raw translation) and APE an automatically post-edited translation.

In the first example in Table 4, the confidence of RBMT result decreases by 1.83 points from 4.8 after automatic post-editing. In fact, the translation sentence after automatic post-editing has gotten worse obviously.

Contrastingly, in the second example, the quality of the translation has improved by 0.13 point from raw translation after automatic post-editing. Notice the translation of the term "バリ取り作業" in the source sentence, translated as 'deburring work' in the reference, changes from 'barricade picking work' to 'deburring work' after automatic post-editing.

Source / Target sentence		Confidence of RBMT	Diff. of Confidence
SRC	なお、トップ３ｔは、仮想の存在である。	4.80	-1.83
REF	the top 3t has an imaginary existence .		
RAW	3 t of tops are existence of imagination .		
APE	3 t of the frames are presence or absence of virtual .		
SRC	バリ取り作業にロボットを利用することは従来より公知の技術である。	2.52	+0.13
REF	the use of a robot for deburring work is a known prior art .		
RAW	it is technology better known than before to use a robot for barricade picking work .		
APE	it is better known than before to use a robot for deburring work .		

Table 4 : Example and the difference of confidence between raw translation and automatically post-edited translation.

The proposed method is more robust than previous methods. For one thing, it might be able to properly treat 'out-of-the-blue' words considered as one of the inherent problems in SMT-based APE, since post-editing is applied selectively according to the quality prediction.

However, the problem of choosing the appropriate confidence threshold remains unsolved. (Specia, Saunders, Turchi, Wang, & Shawe-Taylor, 2009) suggest to find an appropriate threshold by binary search under some significance level. The system in this paper will also need a similar approach to select the threshold.

Another limitation is that I only used the translational information for the feature set for PLS regression analysis. For further improvement in prediction accuracy, information in source sentences should also be used, which could be obtained by word alignment.

References

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., et al. (2003). Confidence estimation for machine translation. *Technical report, Johns Hopkins Univ.*

Grinberg, D., Lafferty, J., & Sleator, D. (1999). A robust parsing algorithm for link grammars. *Proceedings of the 4th International Workshop on Parsing Technologies* .

Koehn, P., Federico, M., Cowan, B., Zens, R., Dyer, C., Bojar, O., et al. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. *Proceedings of the ACL 2007 Demo and Poster Sessions* , 177-180.

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical Phrase-based Translation. *Proceedings of NAACL HLT 2003* , 127-133.

Lagarda, A.-L., Alabau, V., Casacuberta, F., Silva, R., & Diaz-de-Liano, E. (2009, June). Statistical Post-Editing of a Rule-Based Machine Translation System. *Proceedings of NAACL HLT 2009, ACL* , 217-220.

NII:The 7th NTCIR Workshop. (2007/2008). (NII) Retrieved from http://research.nii.ac.jp/ntcir/ntcir-ws7/ws-en.html

NIST: Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurence Statistics. (2002). Retrieved from ftp://jaguar.ncsl.nist.gov/mt/mt2001/mt-eval-02-jan-public.pdf

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Tree. *Proceedings of International Conference on New Methods in Language Processing* .

Simard, M., Goutte, C., & Isabelle, P. (2007, April). Statistical Phrase-based Post-editing. *Proceedings of NAACL HLT 2007, ACL* , 508-515.

Simard, M., Ueffing, N., Isabelle, P., & Kuhn, R. (2007, June). Rule-based Translation With Statistical Phrase-based Post-editing. *Proceedings of the second Workshop on Statistical Machine Translation, ACL* , 203-206.

Specia, L., Cancedda, N., Turchi, M., & Cristianini, N. (2009, May). Estimating the Sentence-Level Quality of Machine Translation Systems. *Proceedings of the 13th Annual Conference of the EAMT* , 28-35.

Specia, L., Saunders, C., Turchi, M., Wang, Z., & Shawe-Taylor, J. (2009). Improving the Confidence of Machine Translation Quality Estimates. *MT Summit XII* .

Wold, S., Ruhe, A., Wold, H., & Dunn, W. J. (1984). The covariance problem in linear regression. the partial least squares(pls) approach to generalized inverses. *SIAM Journal on Scientific Computing , 5*, 735-743.

Zhu, X., Yang, M., Wang, L., Wang, J., & Li, S. (2009). A Quantitative Analysis of Linguistic Factors in Human Translation Evaluation. *2nd International Symposium on Knowledge Acquisition and Modeling* , 410-413.

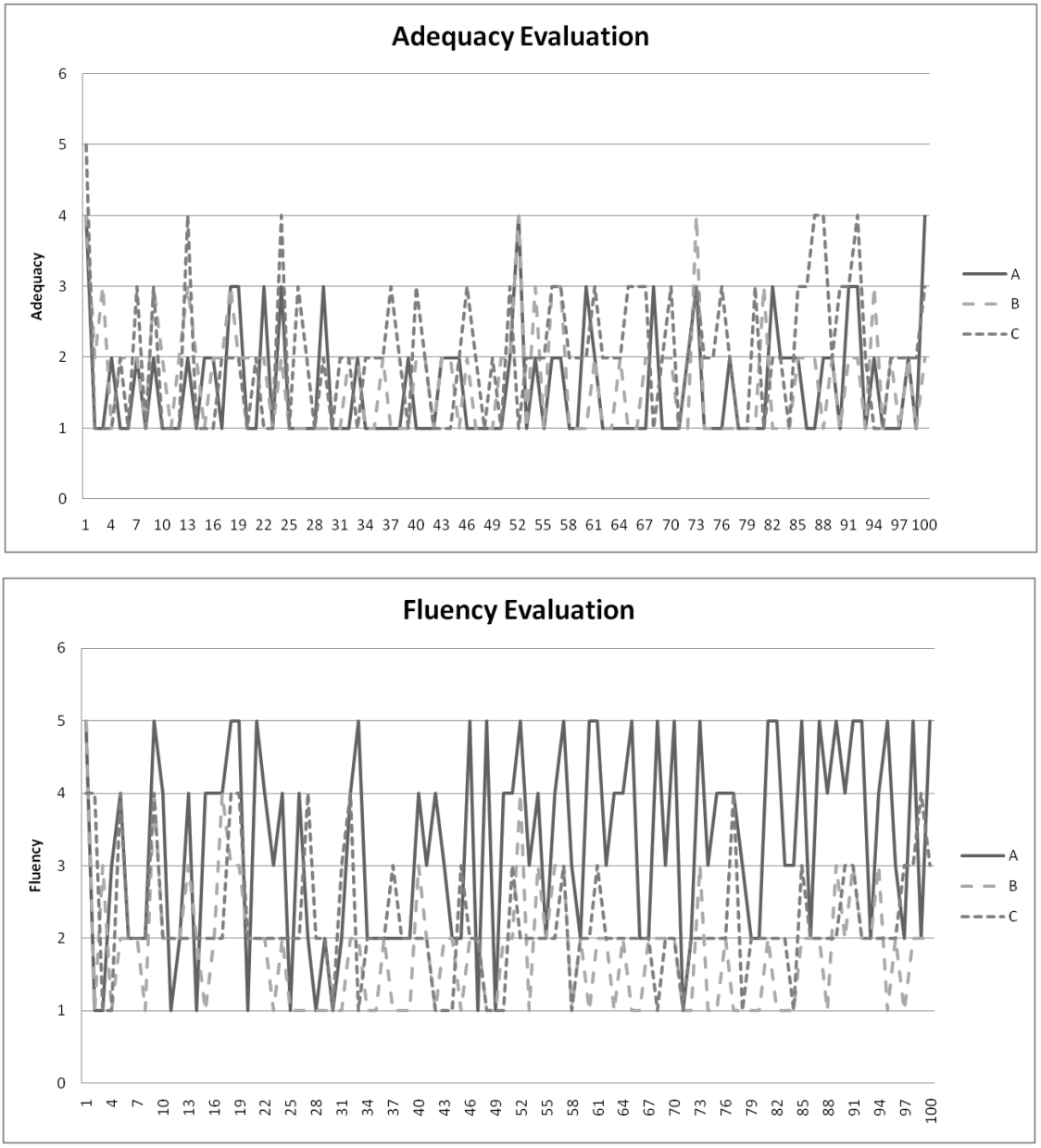

Fig. 5 : Adequacy/Fluency evaluations by 3 human raters for 100 sentences