Adaptive HTER Estimation for Document-Specific MT Post-Editing

Fei Huang *
Facebook Inc.
Menlo Park, CA
feihuang@fb.com

Jian-Ming Xu **Abraham Ittycheriah** **Salim Roukos**
IBM T.J. Watson Research Center
Yorktown Heights, NY
{jianxu, abei, roukos}@us.ibm.com

Abstract

We present an adaptive translation quality estimation (QE) method to predict the human-targeted translation error rate (HTER) for a document-specific machine translation model. We first introduce features derived internal to the translation decoding process as well as externally from the source sentence analysis. We show the effectiveness of such features in both classification and regression of MT quality. By dynamically training the QE model for the document-specific MT model, we are able to achieve consistency and prediction quality across multiple documents, demonstrated by the higher correlation coefficient and F-scores in finding *Good* sentences. Additionally, the proposed method is applied to IBM English-to-Japanese MT post editing field study and we observe strong correlation with human preference, with a 10% increase in human translators' productivity.

1 Introduction

Machine translation (MT) systems suffer from an inconsistent and unstable translation quality. Depending on the difficulty of the input sentences (sentence length, OOV words, complex sentence structures and the coverage of the MT system's training data), some translation outputs can be perfect, while others are ungrammatical, missing important words or even totally garbled. As a result, users do not know whether they can trust the translation output unless they spend time to analyze the MT output. This shortcoming is one of the main obstacles for the adoption of MT systems, especially in machine assisted human translation: MT post-editing, where human translators have an option to edit MT proposals or translate from scratch. It has been observed that human translators often discard MT proposals even if some are very accurate. If MT proposals are used properly, post-editing can increase translators productivity and lead to significant cost savings. Therefore, it is beneficial to provide MT confidence estimation, to help the translators to decide whether to accept MT proposals, making minor modifications on MT proposals when the quality is high or translating from scratching when the quality is low. This will save the time of reading and parsing low quality MT and improve user experience.

In this paper we propose an adaptive quality estimation that predicts sentence-level human-targeted translation error rate (HTER) (Snover et al., 2006) for a document-specific MT post-editing system. HTER is an ideal quality measurement for MT post editing since the reference is obtained from human correction of the MT output. Document-specific MT model is an MT model that is specifically built for the given input document. It is demonstrated in (Roukos et al., 2012) that document-specific MT models significantly improve the translation quality. However, this raises two issues for quality estimation. First, existing approaches to MT quality estimation rely on lexical and syntactical features defined over parallel sentence pairs, which includes source sentences, MT outputs and references, and translation models (Blatz et al., 2004; Ueffing and Ney, 2007; Specia et al., 2009a; Xiong et al., 2010; Soricut and Echihabi, 2010a; Bach et al., 2011). Therefore, when the MT quality estimation model is trained,

*This work was done when the author was with IBM Research.

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 861–870,
Baltimore, Maryland, USA, June 23-25 2014.

it can not be adapted to provide accurate estimates on the outputs of document-specific MT models. Second, the MT quality estimation might be inconsistent across different document-specific MT models, thus the confidence score is unreliable and not very helpful to users.

In contrast to traditional static MT quality estimation methods, our approach not only trains the MT quality estimator dynamically for each document-specific MT model to obtain higher prediction accuracy, but also achieves consistency over different document-specific MT models. The experiments show that our MT quality estimation is highly correlated with human judgment and helps translators to increase the MT proposal adoption rate in post-editing.

We will review related work on MT quality estimation in section 2. In section 3 we will introduce the document-specific MT system built for post-editing. We describe the static quality estimation method in section 4, and propose the adaptive quality estimation method in section 5. In section 6 we demonstrate the improvement of MT quality estimation with our method, followed by discussion and conclusion in section 7.

2 Related Work

There has been a long history of study in confidence estimation of machine translation. The work of (Blatz et al., 2004) is among the best known study of sentence and word level features for translation error prediction. Along this line of research, improvements can be obtained by incorporating more features as shown in (Quirk, 2004; Sanchis et al., 2007; Raybaud et al., 2009; Specia et al., 2009b). Soricut and Echihabi (2010b) proposed various regression models to predict the expected BLEU score of a given sentence translation hypothesis. Ueffing and Hey (2007) introduced word posterior probabilities (WPP) features and applied them in the n-best list reranking. Target part-of-speech and null dependency link are exploited in a MaxEnt classifier to improve the MT quality estimation (Xiong et al., 2010).

Quality estimation focusing on MT post-editing has been an active research topic, especially after the WMT 2012 (Callison-Burch et al., 2012) and WMT2013 (Bojar et al., 2013) workshops with the "Quality Estimation" shared task. Biçici et al. (2013) proposes a number of features measuring the similarity of the source sentence to the source side of the MT training corpus, which, combined with features from translation output, achieved significantly superior performance in the MT QE evaluation. Felice and Specia (2012) investigates the impact of a large set of linguistically inspired features on quality estimation accuracy, which are not able to outperform the shallower features based on word statistics. González-Rubio et al. (2013) proposed a principled method for performing regression for quality estimation using dimensionality reduction techniques based on partial least squares regression. Given the feature redundancy in MT QE, their approach is able to improve prediction accuracy while significantly reducing the size of the feature sets.

3 Document-specific MT System

In our MT post-editing setup, we are given documents in the domain of software manuals, technical outlook or customer support materials. Each translation request comes as a document with several thousand sentences, focusing on a specific topic, such as the user manual of some software.

The input documents are automatically segmented into sentences, which are also called *segments*. Thus in the rest of the paper we will use sentences and segments interchangeably. Our parallel corpora includes tens of millions of sentence pairs covering a wide range of topics. Building a general MT system using all the parallel data not only produces a huge translation model (unless with very aggressive pruning), the performance on the given input document is suboptimal due to the unwanted dominance of out-of-domain data. Past research suggests using weighted sentences or corpora for domain adaptation (Lu et al., 2007; Matsoukas et al., 2009; Foster et al., 2010). Here we adopt the same strategy, building a document-specific translation model for each input document.

The document-specific system is built based on sub-sampling: from the parallel corpora we select sentence pairs that are the most similar to the sentences from the input document, then build the MT system with the sub-sampled sentence pairs. The *similarity* is defined as the number of n-grams that appear in both source sentences, divided by the input sentence's length, with higher weights assigned to longer n-grams. From the extracted sentence pairs, we utilize the standard pipeline in SMT system building: word align-

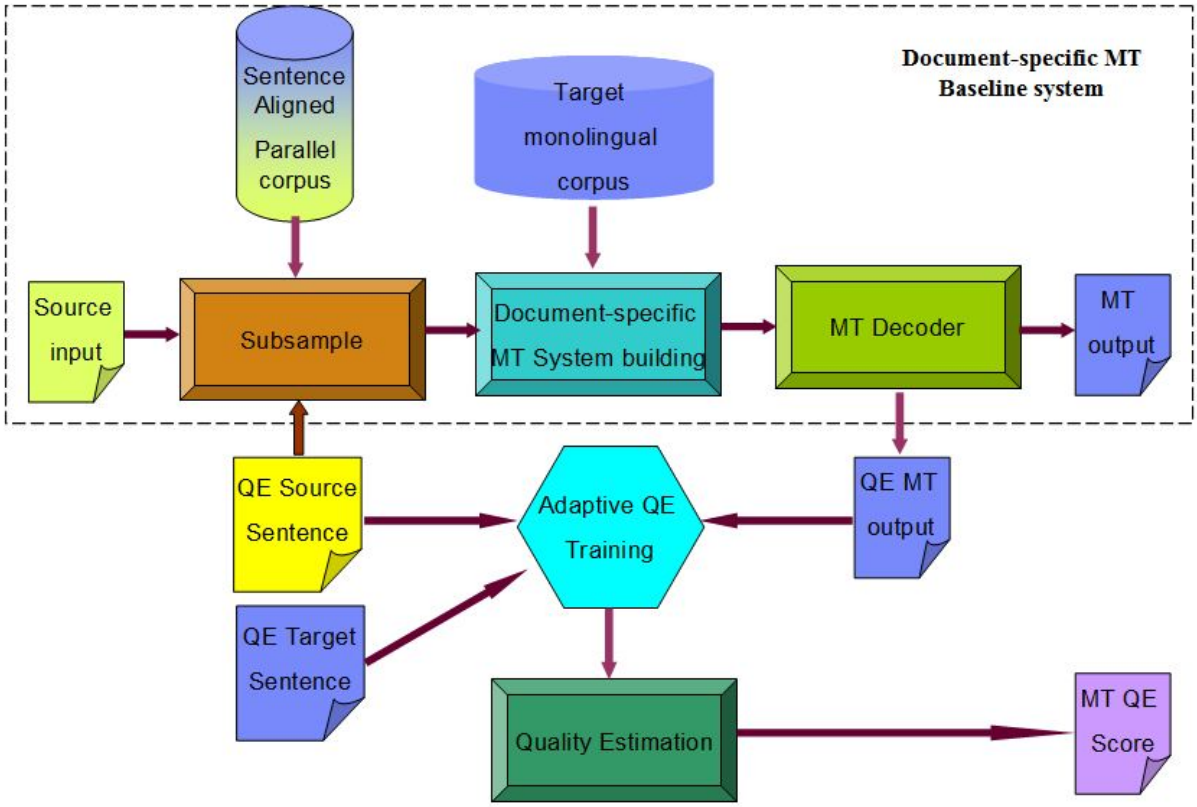

Figure 1: Adaptive QE for document-specific MT system.

ment (HMM (Vogel et al., 1996) and MaxEnt (Ittycheriah and Roukos, 2005) alignment models, phrase pair extraction, MT model training (Ittycheriah and Roukos, 2007) and LM model training. The top region within the dashed line in Figure 1 shows the overall system built pipeline.

3.1 MT Decoder

The MT decoder (Ittycheriah and Roukos, 2007) employed in our study extracts various features (source words, morphemes and POS tags, target words and POS tags, etc.) with their weights trained in a maximum entropy framework. These features are combined with other features used in a typical phrase-based translation system. Altogether the decoder incorporates 17 features with weights estimated by PRO (Hopkins and May, 2011) in the decoding process, and achieves state-of-the-art translation performance in various Arabic-English translation evaluations (NIST MT2008, GALE and BOLT projects).

4 Static MT Quality Estimation

MT quality estimation is typically formulated as a prediction problem: estimating the confidence score or translation error rate of the translated sentences or documents based on a set of features. In this work, we adopt HTER in (Snover et al., 2006) as our prediction output. HTER measures the percentage of insertions, deletions, substitutions and shifts needed to correct the MT outputs. In the rest of the paper, we use TER and HTER interchangably.

In this section we will first introduce the set of features, and then discuss MT QE problem from classification and regression point of views.

4.1 Features for MT QE

The features for quality estimation should reflect the complexity of the source sentence and the decoding process. Therefore we conduct syntactic analysis on the source sentences, extract features from the decoding process and select the following 26 features:

- 17 decoding features, including phrase translation probabilities (source-to-target and target-to-source), word translation probabilities (also in both directions), maxent probabilities[1], word count, phrase count, distor-

[1]The maxent probability is the translation probability

tion probabilities, as well as a set of language model scores.

- Sentence length, i.e., the number of words in the source sentence.
- Source sentence syntactic features, including the number of *noun phrases, verb phrases, adjective phrases, adverb phrases*, as inspired by (Green et al., 2013).
- The length of verb phrases, because verbs are typically the roots in dependency structure and they have more varieties during translation.
- The maximum length of source phrases in the final translation, since longer matching source phrase indicates better coverage of the input sentence with possibly better translations.
- The number of phrase pairs with high *fuzzy match (FM)* score. The high FM phrases are selected from sentence pairs which are closest in terms of n-gram overlap to the input sentence. These sentences are often found in previous translations of the software manual, and thus are very helpful for translating the current sentence.
- The average translation probability of the phrase translation pairs in the final translation, which provides the overall translation quality on the phrase level.

The first 17 features come from the decoding process, which are called "decoding features". The remaining 9 features not related to the decoder are called "external features". To evaluate the effectiveness of the proposed features, we train various classifiers with different feature configurations to predict whether a translation output is useful (with lower TER) as described in the following section.

4.2 MT QE as Classification

Predicting TER with various input features can be treated as a regression problem. However for the post-editing task, we argue that it could also be cast as a classification problem: MT system users (including the translators) are often interested to know whether a given translation is reasonably good or not. If useful, they can quickly look through the translation and make minor modifications. On the other hand, they will just skip reading and parsing the bad translation, and prefer to translate by themselves from scratch. Therefore we also develop algorithms that classify the translation at different levels, depending on whether the TER is less than a given threshold. In our experiments, we set TER=0.1 as the threshold.

Configuration	Training set	Test set
Baseline (All negative)	80%	77%
17 decoding features only	89%	79%
9 external features only	85%	81%
total 26 features	92%	83%

Table 1: QE classification accuracy with different feature configurations

We randomly select one input document with 2067 sentences for the experiment. We build a document-specific MT system to translate this document, then ask human translator to correct the translation output. We compute TER for each sentence using the human correction as the reference. The TER of the whole document is 0.31, which means about 30% errors should be corrected. In the classification task, our goal is to predict whether a sentence is a *Good* translation (with TER $\leq$ 0.1), and label them for human correction. We adopt a decision tree-based classifier, experimenting with different feature configurations. We select the top 1867 sentences for training and the bottom 200 sentences for test. In the test set, there are 46 sentences with TER $\leq$ 0.1. Table 1 shows the classification accuracy.

First we can see that as the overall TER is around 0.3, predicting all the sentences being *negative* already has a strong baseline: 77%. However this is not helpful for the human translators, because that means they have to translate every sentence from scratch, and consequently there is no productivity gain from MT post-editing. If we only use the 17 decoding features, it improves the classification accuracy by 9% on the training set, but only 2% on the test set. This is probably due to the overfitting when training the decision tree classifier. While using the 7 external features, the gain on training set is less but the gain on the test set

derived from a Maximum Entropy translation model (Ittycheriah and Roukos, 2005).

is greater (4% improvement), because the translation output is generated based on the log-linear combination of these decoding features, which are biased towards the final translations. The external features capture the syntactic structure of the source sentence, as well as the coverage of the training data with regard to the input sentence, which are good indicators of the translation quality. Combining both the decoding features and the external features, we observed the best accuracy on both the training and test set. We will use the combined 26 features in the following work.

4.3 MT QE as Regression

For the QE regression task, we predict the TER for each sentence translation using the above 26 features. We experiment with several classifiers: linear regression model, decision tree based regression model and SVM model. With the same training and test data set up, we predict the TER for each sentence in the test set, and compute the correlation coefficient (r) and root mean square error (RMSE). Our experiments show that the decision tree-based regression model obtains the highest correlation coefficients (0.53) and lowest RMSE (0.23) in both the training and test sets. We will use this model for the adaptive MT QE in the following work.

5 Adaptive MT Quality Estimation

The above QE regression model is trained on a portion of the sentences from the input document, and evaluated on the remaining sentences from the same document. One would like to know whether the trained model can achieve consistent TER prediction accuracy on other documents. When we use the cross-document models for prediction, the correlation is significantly worse (the details are discussed in section 6.1). Therefore it is necessary to build a QE regression model that's robust to different document-specific translation models. To deal with this problem, we propose this adaptive MT QE method described below.

Our proposed method is as follows: we select a fixed set of sentence pairs (S_q, R_q) to train the QE model. The source side of the QE training data S_q is combined with the input document S_d for MT system training data subsampling. Once the document-specific MT system is trained, we use it to translate both the input document and the source QE training data, obtaining the translation T_d and

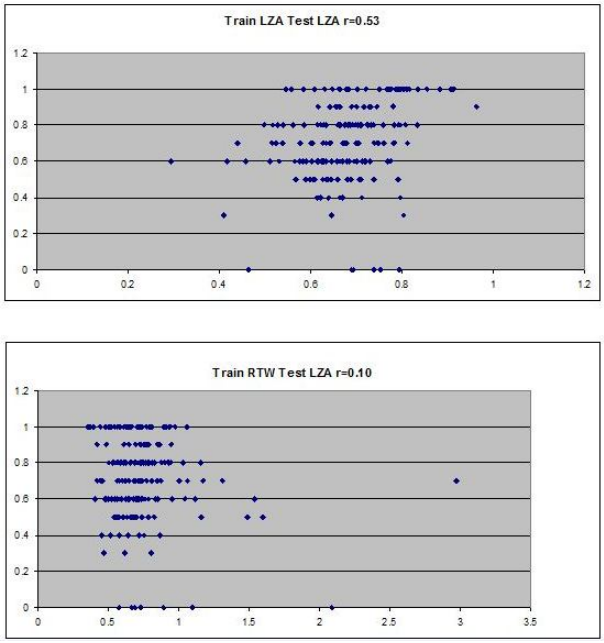

Figure 2: Correlation coefficient r between predicted TER (x-axis) and true TER (y-axis) for QE models trained from the same document (top figure) or different document (bottom figure).

T_q . We compute the TER of T_q using R_q as the reference, and train a QE regression model with the 26 features proposed in section 4.1. Then we use this *document-specific* QE model to predict the TER of the document translation T_d. As the QE model is adaptively re-trained for each document-specific MT system, its prediction is more accurate and consistent. Figure 1 shows the flow of our MT system with the adaptive QE training integrated as part of the built.

6 Experiments

In this section, we first discuss experiments that compare adaptive QE method and static QE method on a few documents, and then present results we obtained after deploying the adaptive QE method in an English-to-Japanese MT Post-Editing project. As mentioned before, the main motivation for us to develop MT QE classification scheme is that translators often discard good MT proposals and translate the segments from scratch. We would like to provide translators with some guidance on reasonably good MT proposals–the sentences with low TERs–to help them increase the leverage on MT proposals to achieve improved productivity.

6.1 Evaluation on Test Set

Our experiment and evaluation is conducted over three documents, each with about 2000 segments. We first build document-specific MT model for each document, then ask human translators to correct the MT outputs and obtain the reference translation. In a typical MT QE scenario, the QE model is pre-trained and applied to various MT outputs, even though the QE training data and MT outputs are generated from different translation models. To evaluate whether such model mismatch matters, we compare the cross-model QE with the same-model QE, where the QE training data and the MT outputs are generated from the same MT model.

We select one document LZA with 2067 sentences. We use the first 1867 sentences to train the static QE model and the remaining 200 sentences are used as test set for TER prediction. We compute the correlation coefficient (r) between each predicted TER and true TER, as shown in Figure 2. We find that the TER predictions are reasonably correct when the training and test sentences are from the same MT model (the top figure), with correlation coefficients around 0.5. For the cross-model QE, we train a static QE model with 1867 sentences from another document RTW, and use it to predict the TER of the same 200 sentences from document LZA (the bottom figure). We observe significant degradation of correlation coefficient, dropping from 0.5 to 0.1. This degradation and unstable nature is the prime motivation to develop a more robust MT quality estimation model.

We select 1700 sentences from multiple previously translated documents as the QE training data, which are independent of the test documents. We train the static QE model with this training set, including the source sentences, references and MT outputs (from multiple translation models). To train the adaptive QE model for each test document, we build a translation model whose subsampling data includes source sentences from both the test document and the QE training data. We translate the QE source sentences with this newly built MT model, and the translation output is used to train the QE model specific to each test document. We compare these two QE models on three documents, LZA, RTW and WC7, measuring r and RMSE for each QE model. The result is shown in Table 2. We find that the adaptive QE model demonstrates higher r and lower RMSE than the static QE model for all the test documents.

Besides the general correlation with human judgment, we particularly focus on those reasonably good translations, i.e., the sentences with low TERs which can help improve the translator's productivity most. Here we report the precision, recall and F-score of finding such "*Good*" sentences (with TER ≤ 0.1) on the three documents in Table 3. Again, the adaptive QE model produces higher recall, mostly higher precision, and significantly improved F-score. The overall F-score of the adaptive QE model is 0.28[2]. Compared with the static QE model's 0.17 F-score, this is relatively 64% improvement.

In the adaptive QE model, the source side QE training data is included in the subsampling process to build the document-specific MT model. It would be interesting to know whether this process will negatively affect the MT quality. We evaluate the TER of MT outputs with and without the adaptive QE training on the same three documents. As seen in Table 4, we do not notice translation quality degradation. Instead, we observe slightly improvement on two document, with TERs reduction by 0.1-0.4 pt. As our MT model training data include proprietary data, the MT performance is significantly better than publicly available MT software.

6.2 Impact on Human Translators

We apply the proposed adaptive QE model to large scale English-to-Japanese MT Post-Editing project on 36 documents with 562K words. Each English sentence can be categorized into 3 classes:

- Exact Match (**EM**): the source sentence is completely covered in the bilingual training corpora thus the corresponding target sentence is returned as the translation;
- Fuzzy Match (**FM**): the source sentence is similar to some sentence in the training data (similarity measured by string editing distance), the corresponding fuzzy match target sentence (FM proposal) as well as the MT translation output (MT proposal) are returned for human translators to select and correct;
- No Proposal (**NP**): there is no close match source sentences in the training data (the FM

[2]The adaptive QE model obtains much higher F-score (80%) on the rest of the sentences (with TER > 0.1).

Document	LZA		RTW		WC7	
Num. of Sents	2067		2003		2405	
	$r \uparrow$	RMSE $\downarrow$	$r \uparrow$	RMSE $\downarrow$	$r \uparrow$	RMSE $\downarrow$
Static QE	0.10	0.38	0.40	0.32	0.13	0.36
Adaptive QE	**0.58**	0.23	**0.61**	0.22	**0.47**	0.20

Table 2: QE regression with static and adaptive models

Document	LZA	RTW	WC7
Num. of Sents	2067	2003	2405
	P/R/F-score	P/R/F-score	P/R/F-score
Static QE	0.73/0.08/0.14	0.69/ 0.11/ 0.19	0.74/ 0.10/ 0.18
Adaptive QE	0.69/0.14/**0.24**	0.84/ 0.16/ **0.26**	0.80/ 0.23/ **0.35**

Table 3: Performance on predicting *Good* sentences with static and adaptive models

similarity score of 70% is used as the threshold), therefore only the MT output is returned.

EM sentences are excluded from the study because in general they do not require editing. We focus on the FM and NP sentences[3]. In Table 5 we present the precision, recall and F-score of the "Good" sentences in the FM and NP categories, similar to those shown in Table 3. We consistently observe higher performance on the FM sentences, in terms of precision, recall and F-score. This is expected because these sentences are well covered in the training data. The overall F-score is in line with the test set results shown in Table 3.

We are also interested to know whether the proposed adaptive QE method is helpful to human translators in the MT post-editing task. Based on the TERs predicted by the adaptive QE model, we assign each MT proposal with a confidence label: High ($0 \leq$ TER ≤ 0.2), Medium ($0.2 <$ TER ≤ 0.3), or Low (TER > 0.3). We present the MT proposals with confidence labels to human translators, then measure the percentage of sentences whose MT proposals are used. From Table 6 and 7, we can see that sentences with High and Medium confidence labels are more frequently used by the translators than those with Low labels, for both the FM and NP categories. The MT usage for the FM category is less than that for the NP category because translators can choose FM proposals instead of the MT proposals for correction.

We also measure the translator's productivity gain for MT proposals with different confidence labels. The productivity of a translator is defined as the number of source words translated per unit time. The post editing tool, IBM TranslationManager, records the time that a translator spends on a segment and computes the number of characters that a translator types on the segment so that we can compute how many words the translator has finished in a given time.

[3]The word count distribution of EM, FM and NP is 21%, 38% and 41%, respectively.

Document	LZA	RTW	WC7
TER-Baseline	30.81	30.74	29.96
TER-with Adaptive QE	30.69	30.78	29.56

Table 4: MT Quality with and without Adaptive QE measured by TER

We choose the overall productivity of NP0 as the base unit 1, where there is no proposal presents and the translator has to translate the segments from scratch. Measured with this unit, for example, the overall productivity of FM0 being 1.14 implies a relative gain of 14% over that of NP0, which demonstrates the effectiveness of FM proposals.

Table 6 and 7 also show the productivity gain on sentences with High, Medium and Low labels from FM and NP categories. Again, the productivity gain is consistent with the confidence labels from the adaptive QE model's prediction. The overall productivity gain with confidence-labeled MT proposals is about 10% (comparing FM1 vs. FM0 and NP1 vs. NP0). These results clearly demonstrate the effectiveness of the adaptive QE model in aiding the translators to make use of MT proposals and improve productivity.

Category	Class	FM usage	**MT usage**	**Productivity**
FM1	High	33%	**34%**	**1.35**
	Medium	47%	**18%**	**1.21**
	Low	60%	**8%**	**1.20**
	Overall	45%	**21%**	**1.26**
FM0	High	53%	-	1.12
	Medium	64%	-	1.14
	Low	67%	-	1.16
	Overall	59%	-	1.14

Table 6: MT proposal usage and productivity gain in FM category.
In FM1, both Fuzzy Match and MT proposals present. In control class FM0, only Fuzzy Match proposals present, and therefore, MT usage is not available for FM0. Strong correlation is observed between predicted "High" , "Medium" and "Low" sentences with MT usage and post editing productivity.

Category	Class	**MT usage**	**Productivity**
NP1	High	**50%**	**1.25**
	Medium	**42%**	**1.08**
	Low	**27%**	**1.00**
	Overall	**38%**	**1.09**
NP0	High	-	1.08
	Medium	-	1.00
	Low	-	0.96
	Overall	-	1.00

Table 7: MT proposal usage and productivity gain in NP category.
In NP1, MT is the only proposal available, while in control NP0, there presents no proposal at all and the translator has to translate from scratch. Strong correlation is observed between predicted "High" , "Medium" and "Low" sentences with MT usage and post editing productivity

Type	Precision	Recall	F-score
FM	0.71	0.23	0.35
NP	0.67	0.18	0.29
Overall	**0.69**	**0.21**	**0.32**

Table 5: Performance on predicting *Good* sentences (TER ≤ 0.1) by adaptive QE model

7 Discussion and Conclusion

In this paper we proposed a method to adaptively train a quality estimation model for document-specific MT post editing. With the 26 proposed features derived from decoding process and source sentence syntactic analysis, the proposed QE model achieved better TER prediction, higher correlation with human correction of MT output and higher F-score in finding good translations. The proposed adaptive QE model is deployed to a large scale English-to-Japanese MT post editing project, showing strong correlation with human preference and leading to about 10% gain in human translator productivity.

The training data for QE model can be selected independent of the input document. With such fixed QE training data, it is possible to measure the consistency of the trained QE models, and to allow the sanity check of the document-specific MT models. However, adding such data in the subsampling process extracts more bilingual data for building the MT models, which slightly increase the model building time but increased the translation quality. Another option is to select the sentence pairs from the MT system subsampled training data, which is more similar to the input document thus the trained QE model could be a better match to the input document. However, the QE model training data is no longer constant. The model consistency is no longer guaranteed, and the QE training data must be removed from the MT system training data to avoid data contamination.

References

Nguyen Bach, Fei Huang, and Yaser Al-Onaizan. 2011. Goodness: A method for measuring machine translation confidence. In *ACL*, pages 211–219.

Ergun Biçici, Declan Groves, and Josef van Genabith. 2013. Predicting sentence translation quality using extrinsic and language independent features. *Machine Translation*.

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.

Ondrej 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 *Eighth Workshop on Statistical Machine Translation*, WMT-2013, pages 1–44, Sofia, Bulgaria.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada.

Mariano Felice and Lucia Specia. 2012. Linguistic features for quality estimation. In *Seventh Workshop on Statistical Machine Translation*, pages 96–103, Montréal, Canada.

George Foster, Cyril Goutte, and Roland Kuhn. 2010. Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 451–459, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jesús González-Rubio, Jose Ramón Navarro-Cerdán, and Francisco Casacuberta. 2013. Dimensionality reduction methods for machine translation quality estimation. *Machine Translation*, 27(3-4):281–301.

Spence Green, Jeffrey Heer, and Christopher D. Manning. 2013. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 439–448, New York, NY, USA. ACM.

Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1352–1362, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Abraham Ittycheriah and Salim Roukos. 2005. A maximum entropy word aligner for arabic-english machine translation. In *In Proceedings of HLT-EMNLP*, pages 89–96.

Abraham Ittycheriah and Salim Roukos. 2007. Direct translation model 2. In *In HLT-NAACL 2007: Main Conference*, pages 57–64.

Yajuan Lu, Jin Huang, and Qun Liu. 2007. Improving statistical machine translation performance by

training data selection and optimization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 343–350, Prague, Czech Republic, June. Association for Computational Linguistics.

Spyros Matsoukas, Antti-Veikko I. Rosti, and Bing Zhang. 2009. Discriminative corpus weight estimation for machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 708–717, Stroudsburg, PA, USA. Association for Computational Linguistics.

Christopher B. Quirk. 2004. Training a sentence-level machine translation confidence measure. In *In Proceedings of LREC*.

Sylvain Raybaud, Caroline Lavecchia, David Langlois, and Kamel Smaïli. 2009. New confidence measures for statistical machine translation. *CoRR*, abs/0902.1033.

Salim Roukos, Abraham Ittycheriah, and Jian-Ming Xu. 2012. Document-specific statistical machine translation for improving human translation productivity. In *Proceedings of the 13th international conference on Computational Linguistics and Intelligent Text Processing - Volume Part II*, CICLing'12, pages 25–39, Berlin, Heidelberg. Springer-Verlag.

Alberto Sanchis, Alfons Juan, Enrique Vidal, and Departament De Sistemes Informtics. 2007. Estimation of confidence measures for machine translation. In *In Procedings of Machine Translation Summit XI*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Radu Soricut and Abdessamad Echihabi. 2010a. Trustrank: inducing trust in automatic translations via ranking. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 612–621, Stroudsburg, PA, USA. Association for Computational Linguistics.

Radu Soricut and Abdessamad Echihabi. 2010b. Trustrank: Inducing trust in automatic translations via ranking. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 612–621. Association for Computational Linguistics.

Lucia Specia, Craig Saunders, Marco Turchi, Zhuoran Wang, and John Shawe-taylor. 2009a. Improving the confidence of machine translation quality estimates. In *In Proceedings of MT Summit XII*.

Lucia Specia, Marco Turchi, Zhuoran Wang, John Shawe-Taylor, and Craig Saunders. 2009b. Improving the confidence of machine translation quality estimates.

Nicola Ueffing and Hermann Ney. 2007. Word-level confidence estimation for machine translation. *Computational Linguistics*, 33(1):9–40.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. Hmm-based word alignment in statistical translation. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, COLING '96, pages 836–841, Stroudsburg, PA, USA. Association for Computational Linguistics.

Deyi Xiong, Min Zhang, and Haizhou Li. 2010. Error detection for statistical machine translation using linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 604–611, Stroudsburg, PA, USA. Association for Computational Linguistics.