MoBiL: A Hybrid Feature Set for Automatic Human Translation Quality Assessment

Yu Yuan[1], **Serge Sharoff**[2], **Bogdan Babych**[3]
Nanjing University of Information Science &Technology[1]
University of Leeds[2,3]
LS2 9JT
E-mail: hittle.yuan@gmail.com; {s.sharoff; b.babych}@leeds.ac.uk

Abstract

In this paper we introduce **MoBiL**, a hybrid **Mo**nolingual, **Bi**lingual and **L**anguage modelling feature set and feature selection and evaluation framework. The set includes translation quality indicators that can be utilized to automatically predict the quality of human translations in terms of content adequacy and language fluency. We compare **MoBiL** with the QuEst baseline set by using them in classifiers trained with support vector machine and relevance vector machine learning algorithms on the same data set. We also report an experiment on feature selection to opt for fewer but more informative features from **MoBiL**. Our experiments show that classifiers trained on our feature set perform consistently better in predicting both adequacy and fluency than the classifiers trained on the baseline feature set. **MoBiL** also performs well when used with both support vector machine and relevance vector machine algorithms.

Keywords: Translation Quality, Feature Selection, Text Classification, Machine Learning

1. Introduction

Progress in a number of Natural Language Processing and Machine Learning technologies in the past 40 years led to the development of applications for automated scoring of essays and language testing. Systems of automated essay scoring have been sufficiently reliable to be commercialized and deployed for large international language test (Dodigovic, 2005: 104). Some well-known systems include Project Essay Grader, E2rater, Intellimetric and PaperRater, to just name a few, but it is not the case with human translation assessment, particularly student translations.

A different set of technologies have been applied for automated translation quality evaluation, where a lot of attention has been paid to automatic evaluation of Machine Translation (MT), in the form of methods based on parallel corpora, for example, BLEU (Papineni et al., 2002), or MT Quality Estimation (QE), which estimates the suitability of MT output without a reference translation, QuEst (Shah et al., 2013) selects high quality MT translations (Ma & McKeown, 2013) or detects machine translation errors (Xiong, Zhang, & Li, 2010). However, automatic Translation Quality Assessment (TQA) for human translations is a more complicated problem compared to the automated evaluation of MT. One of the reasons is that machine translations, generally inferior to human translation, usually contain a much smaller range of translation errors in comparison with human translations. Human translation errors only partially overlap with errors made by MT and display greater variability, which makes human translations less predictable, see section 4.2 our pilot experiment for reference. As a result, this non-uniformity of the creative human choices and translations further complicates the task of automated estimation of human translation quality, including automated scoring of trainee translators' work, and so far this area remains under-researched.

Useful research in this area comes from the language learning and translator training scenarios, where some attempts have been made to automatically evaluate students' translations on a large scale (Jiang, 2013；Wang & Chang, 2009; Wen, Qin, & Jiang, 2009; Tian, 2008). These experiments treat translation evaluation as an analytical scoring process that takes into account human pre-defined scoring points contained in the translations; they compare and compute similarity of translations in question with reference translations or expert translations. These systems require human annotation of pre-selected scoring points and count the matches of them afterwards. In this respect these approaches are indeed pseudo-automatic, which means that these systems are less likely to be reusable without extensive human intervention into the marking process.

Translation quality assessment (TQA) for human translations differs from evaluation of MT output in several respects. Firstly, as we indicated above, human translations have great variation of otherwise more or less acceptable translation options, even for the same translator, while MT errors are often system-bound, with more or less foreseeable changes. Secondly, MT output is often below what is acceptable, for example, for post-editing tasks, while human translations usually have reasonable quality, at least acceptable for post-editing. Therefore, TQA distinguishes between translations of usable quality that would in any case require only a minimum amount of post-editing, while MT evaluation aims at distinguishing low-quality from reasonable quality MT output (Babych & Hartley, 2008). As the authors propose, automated MT evaluation metrics that compute proximity of MT output to its human gold-standard reference measure structural matches and heavily rely on the lexical level, so they tend to be insensitive to higher-level errors that are more typical for better translations. Human translations typically contain errors beyond the lexical level, to which proximity-based MT evaluation metrics are less sensitive.

In spite of such gaps, automatic MT estimation can still lend some insight into automatic human translation estimation. As in automatic MT evaluation, automatic human TQA can be based on word-level, sentence-level and document-level features, and be formulated as a machine learning task to induce a quality prediction model from labelled or partially labelled data. The whole process of automatic human TQA can be viewed as a text classification task in terms of two main variables: feature engineering and machine learning methods. Motivated by the two factors, we are interested in what performance-based features that better capture the cumulative functional effect of translations at different language levels and are sensitive to relatively high quality human translations.

This paper investigates the plausibility of a set of easily extractable features to estimate the quality of human translations automatically from the viewpoint of adequacy and fluency. We compare the effectiveness of two SVM-based machine learning algorithms, and demonstrate that with feature selection the machine learning models can filter out less relevant features to reduce the complexity and increase the stability of evaluation, while making little sacrifice of prediction quality. The approach has been implemented as an automatic TQA system by which human translation evaluation is treated as a text classification task. Section 2 presents our feature for prediction, including monolingual features (sub section 2.1), bilingual features (sub section 2.2) and language modelling features (sub section 2.2); Section 3 introduces the machine learning methods employed in the research including the baseline system (sub section 3.1), SVM-based classifiers (sub section 3.2 & 3.3) and the feature selection technique (sub section 3.4); Section 4 describes our dataset (sub section 4.1) and reports the findings (sub section 4.2); Section 5 ends with our conclusions and further research plans.

2. Translation Quality Indicators

Translation is subject to a continuous interaction of inner linguistic-textual factors, e.g., language norms and their constrains, and extra-linguistic factors, such as intertextuality, the translation brief, working conditions, translator's competence and translation receptor's perception (House, 2014:2). Translators make different decisions in dealing with this cross-lingual operation. This often result in the systemic difference of the distribution of syntax-semantic features between translations of different quality levels, for example, high scores translations tend to more conform to target language norms by using idiomatic collocations and demonstrate high smoothness in contrast to wrong collocations and inappropriate choices of words that are common to inferior translations. The feature set proposed here is broadly divided into three categories: monolingual, bilingual and language modelling features. It shall be noted that we here concentrate on language-independent features with the aim to allow for more accessible comparisons of qualities of distantly related language pairs such as English and Chinese.

2.1 Monolingual Features

Hereby monolingual features refer to a range of features that have been available in source texts and target texts (TTs). They include:

All together 11 high frequent part-of-speeches (POS) shared between the STs and TTs are selected as POS features in our experiment. We referred to Universal POS-Tagset to match pos tags in both English source text and Chinese translations in order to achieve better comparability (Petrov, Das & McDonald, 2011). As linguistically motivated features, part-of-speech related features are used as baseline features in the WMT Quality Estimation shared-task 2012 (Callison-Burch et al., 2012). For instance, number, percentage and ratio of content words and function words are extracted as linguistic features in Felice & Specia (2012). In the similar vein, POS tags were counted as shallow grammatical matches on both the source and the target (Avramidis, 2012). We have good reason that it might be contributing to the meaning transference from the source texts to the target texts. As a consequence, the researcher included frequencies of these 11 POS-tags (excluding foreign words as it is sparse) in both STs and TTs as a feature group.

In both the source and the target texts. Dependencies have found their way into translation quality prediction (Fox, 2002; Owczarzak et al., 2007; Padó et al., 2009; Shah et al., 2013). In our experiment, we extracted 28 types of target translation dependency structures, and used them in the prediction model. See Table 1 for the detailed list of monolingual features

#	Features
F1-22	2*11 Part of Speech tags in ST & TT.
F23-78	2*28 dependency relations in ST & TT.
F79-80	Number of tokens in ST & TT.
F81-82	Number of types in ST & TT.
F83-84	Number of content words in ST & TT.
F85-86	Number of function words in ST &TT
F87-88	Number of sentences in ST & TT
F89-90	Number of average sentence lengths
F91-100	2*5 phrases in ST & TT
F101-108	2*4 semantic labels in ST & TT

Table 1 Monolingual Features

2.2 Bilingual Features

Bilingual features here refer to a collection of linguistic features that establish a dynamic relationship between the source and target texts. Intuitively, human evaluators rank the translations by examining closely the totality of the core message and features successfully transferred in relation to the source texts. For this reason, it is necessary that quality estimation takes into account features that link the source and target texts together so that its overreliance on feature in target texts can be mitigated. The bilingual features are mainly logarithmized ratio between the corresponding ST and TT features, and the City Distance of the vectors of features of the same group (e.g., ST & TT Type-Token City Distance is distance of two vectors of

Type-Tokens in both ST and TT), which are included in Table 2.

#	Features
F109	TT to ST type log ratio
F110	TT to ST token log ratio
F111	ST & TT Type-Token City Distance (CD)
F112	ST to TT content words log ratio
F113	ST & TT function words log ratio
F114	ST & TT content & function words CD.
F115-142	28 ST & TT dependency log ratios.
F143	ST & TT dependency CD.
F144-148	5 ST &TT phrases log ratios.
F149	ST & TT phrases CD.
F150	ST & TT POS CD.
F151	ST & TT sentence numbers log ratio.
F152	ST & TT sentence length log ratio
F153-156	4 ST & TT semantic role labels log ratios
F157	ST & TT Semantic role labels CD.

Table 2 Bilingual Features

2.3 Language Modelling Features

It is very useful to model a prior distribution over sequences of words and tell which are or are not probable in language. In order to automatically rank the translation produced, a statistical language model of the target language should be first built and then applied to judge the probability and perplexity of the target text. The higher ranking output is deemed to be the more fluent, and therefore better translation.

The above list of LM features are quite self-explanatory, except for the bilingual word embeddings. Hereinafter, we use a few words to introduce the bilingual word embeddings and how it is used in our research. Bilingual embeddings, in its name, refers to the distributed representations of low-dimensional, real-valued vectors for each word across two languages typically induced via neural language models or spectral methods from aligned corpora. It has the advantages for NLP tasks like parsing, sentiment analysis and information retrieval in that this method supplements the labelled data in semi-supervised settings to overcome the inherent data sparsity common to high dimensional NLP domain(Dhillon et al., 2011). Each dimension of the embeddings captures latent information about a combination of syntactic and semantic word properties, and such induced representations can be used as features in a supervised classifier(typically discriminative). In recent years, there has been an increased interest in using semantic embeddings as high-quality semantic features that embody bilingual translation equivalence across languages. The methods and rationale to train bilingual word embeddings have been well explained in Hermann & Blunsom (2014). Following the same approach, in order to train the English-Chinese word embedding model for this study, we make use of a combination of the English-Chinese part of MultiUN (Eisele & Chen, 2010) and UM corpora (Tian et al., 2014) that is roughly about 312 million tokens for English and 289 million for Chinese of ≈11million lines, with misaligned sentences eliminated from both corpora, and BiCVM[1] code to train the bilingual embedding models (100 dimensions in order to save computing power and time) first and then extract the word embeddings for the source texts and all the translations from the English embedding model and Chinese embedding model respectively. For each word in the source text and its translation, we extracted a 100 dimension vector and then sum the total vectors of all words in the source text and its translation and then concatenate them together as 200 features. This bilingual word embedding feature is then used separately, and in combination with the rest of features to predict the translation quality. Details can be seen as follows in Table 3.

#	Features
F158	TT language model(LM) perplexity score
F159	TT number of OOV
F160	TT POS LM perplexity score
F161	TT sentence LM score
F162	TT sentence POS LM score
F163	TT semantic tightness profile score
F164	TT log translation probability
F165	TT bilingual lexicon coverage
F166-365	2*100 ST&TT bilingual word embeddings

Table 3 Language Modelling Features

3. Machine Learning Methods

3.1 Baseline System

Since there is little research towards the automated human translation quality assessment, there are practically no baseline systems available for comparison. In addition, most automated translation quality estimation research is working on European languages and at the sentence level quality estimation. In this case, we build our own base line system with the 277 manually scored translations with the baseline features from QuEst (Shah et al., 2013) ， excluding only the average token length that is incompatible with Chinese. Our baseline is trained with approximate two-thirds of the human annotated translations（≈250）and test on the remaining 27 samples, using SVM-based machine learning algorithms.

3.2 SVM-based Regression

Our feature set overall contains 165 features consisting of the main categories of Part-of-Speech tags, dependency relations, distance measures and language modelling scores that are spanning from the most basic lexico-grammatical features to abstract syntax-semantic relations and even higher textual features. The feature set is far from comprehensive and entirely based on our intuition and previous works by other scholars. It is perhaps this reason that makes the research even valuable to explore the possibility of expanding and verifying some factual findings of automatic human translation evaluation, a newly emerging direction to interdisciplinary researchers. For comparison, we also train with support vector machine learning algorithm on all 165 features. Like the experiment above, we use the same data split for training and post training test. As

[1] https://github.com/karlmoritz/bicvm

mentioned earlier in Section 2.3, we train with the 200 bilingual word embedding features too. The results will be reported below.

3.3 Relevance Vector Machine

While prediction is of vital importance for any automatic scoring model, the model complexity needs to be taken into account, too. Setting accuracy alone is undesirable as it increases the model complexity, which reduces the training set error, and can easily lead to over-fitting and poor generalization. The Relevance Vector Machine (RVM) is a Bayesian model for regression and classification, which uses a functional form similar to SVM (Tipping, 2001). Generally, trained RVM models utilize many fewer basis functions than the corresponding SVM, therefore, leading to better stability outside the current test set, and often they exhibit superior performance on the test set. In order to explore the possibility of building a simpler yet equally efficient model with RVM, we trained the same dataset with relevance vector machine (ten-fold cross validation and the default parameter settings for predicting the content and fluency score) in the Kernlab package.

3.4 Feature Selection

It is our intuition that some features in our data might be in covariance with each other or simply some of them are not informative. With that in mind, we used simulated annealing, a global search method, which makes small random changes to an initial candidate solution is employed to search for an optimal solution. The algorithm fits the model to all predictors backwards and ranks them according to their importance to it. Through many iterations, the top n predictors yielding the best performance are then selected (Kirkpatrick, 1984; Guyon et al., 2002). The experiment was performed using the Caret[2] package in R.

3.5 Label Propagation

Label propagation finds communities in the real, complex networks. This algorithm, in comparison to others, has advantage in its running time, amount of priori information required, with the exception that it produces an aggregate of multiple solution instead. This approach resembles the k-NN nearest neighbours where closer data points tend to have similar labels (Rios & Sharoff, 2015). More detailed explanation of this algorithm can be found in (Zhu & Ghahramani, 2002). In this study, we will use label spreading implemented on scikit-learn to see if we can improve the prediction accuracy with the majority of unlabelled data in our data set. Details will be reported in the following section.

4. Dataset and Results

4.1Training Data

The trainee translation data come from the English-Chinese translations of Parallel Corpus of Chinese EFL Learners (Wen & Wang, 2008). The translations are produced by upper-intermediate level English and Non-English Majors, and thus can be viewed as trainee translators' work. The data has been processed, sentence aligned and annotated with Stanford-Corenlp for English and Chinese (Manning et al., 2014). We extracted 165 features from both Six English source texts spanning from general texts to mildly scientific domain, with ad-hoc python scripts, plus the 200 word embedding features. Each translation text measures approximately 300-400 Chinese characters. Among the 2119 training samples, we manually scored 277 pieces of them in terms of their adequacy and fluency on a scale of 60 points (mean=38.23, interquartile range=7, range=18) for content adequacy and 40 points (mean= 27.84, interquartile range=8, range=22) for language fluency. Two Chinese native annotators, both are University English teachers and research students in Translation Studies, following a scoring scheme of ATA Certification Programme Rubric for Grading (Version 2011), which measures the performance of a translator against four dimensions ranging from content transfer, terminology and style, idiomatic writing and target language conventions, evaluated the translations based on the degree to which learner translators have transferred the meaning completely and followed the rules and conventions of the target language. The inter-annotator agreement is substantial (α =.77 for adequacy and .89 for fluency).

Other than the above corpus of translations, we also used some crawled Chinese fiction, short stories, essays and scientific documents (approximately 253 million tokens) as well as the Chinese Wikipedia Dump (≈150 million tokens)[3] for training the Chinese language model and Chinese Part-of-Speech language model. The resources for training the bilingual embeddings are introduced in section 2.3.

4.2 Results

In this section, we will report the results of the seven experiments described in the foregoing sections. As is mentioned above, the seven experiments are to compare two feature datasets and two learning algorithms. For the relevance vector machine training, we used the 10-fold cross validation and other default parameter settings, while for the support vector machine training we used 10-fold cross validation with some tuning on the development set. Details of the learning parameters can be found below in Table 4 (Prd. = Prediction, A=adequacy, F=Fluency, c=cost, C=Cross Validation, same as below.). Note that experiment 6 is implemented with scikit-learn package and that' s why it has different parameters configuration, and elsewhere uses the "rbfdot" kernel for SVM and RVM learning as with the R package Kernlab's[4].

[2] http://topepo.github.io/caret/index.html

[3] https://dumps.wikimedia.org/zhwiki/20151002/

[4]https://cran.r-project.org/web/packages/kernlab/

#	Prd.	*c*	*ε*	*σ*	C
1	A	1	0.8	0.064413	10
	F	1	0.9	0.067925	10
2	A	5	0.9	0.000121	10
	F	1	0.9	0.000098	10
3	A	5	0.77	0.001009	10
	F	5	0.77	0.001009	10
4	A	10	0.1	0.005073	10
	F	10	0.1	0.005073	10
5	A	10	0.1	1	10
	F	10	0.1	1	10
6	A F	alpha=0.05, gamma=0.01, kernel='rbf', max_iter=500, n_neighbors=7, tol=0.001			
7	A	1	0.9	0.000424	10
	F	1	0.6	0.000519	10

Table 4 Parameters Setting for the Seven Experiments

In the following, Table 5 reports the results obtained from the above experiments (E. = Experiments, Alg.= Algorithm, Vct.= number of support vectors, RMSE= Root Mean Square Error. Note that all fluency prediction RMSEs have been converted to the same scale as Adequacy for comparison.).

#	E.	Prd.	Alg.	Vct.	RMSE
1	QuEst (Baseline)	A	SVM	73	3.84
			RVM	40	7.54
		F	SVM	106	5.94
			RVM	28	9.77
2	Sub Set (165 Features)	A	SVM	206	4.09
			RVM	53	8.27
		F	SVM	212	6.32
			RVM	**26**	9.35
3	Sub Set After Selection	A 52	SVM	209	4.32
			RVM	43	**5.83**
		F 44	SVM	207	6.44
			RVM	**29**	10.00
4	Embeddings (200)	A	SVM	203	**2.26**
			RVM	118	**3.43**
		F	SVM	210	**2.27**
			RVM	100	**3.25**
5	All Features (365)	A	SVM	199	**4.30**
			RVM	42	**5. 37**
		F	SVM	231	**6.69**
			RVM	28	**8.03**
6	All Features label spreading	A	SVM		10.76
		F	SVM		**7.95**
7	All Features after selection	A 100	SVM	221	4.07
			RVM	142	**4.41**
		F 112	SVM	199	**4.69**
			RVM	106	**4.05**

Table 5 Models Trained in Seven Experiments

From top down to bottom up, experiment 1 reports the results for the QuEst baseline system, and in comparison, experiment 2 is the model trained with the 165 sub feature set without bilingual word embeddings, which shows no difference of significance from the baseline, then follows the training of feature selection on the 165 sub feature set (experiment 3), below which using 200 bilingual word embeddings as features is also tested (experiment 4), followed by the implementation of all the 365 features (experiment 5), by which label spreading technique is adopted to see if label propagation contributes to a better model (experiment 6), and finally we implemented feature selection on the full feature set (experiment 7).

As the table shows, on the same training data set and test data, support vector machine model with the baseline feature set performs slightly better but no significant difference from our 165 sub feature set to predict on content and fluency scores (in experiment 2). Feature selection allows us to achieve almost comparable result to the original 165 sub feature set and the Baseline features on support vector learning of both adequacy and fluency (as in experiment 3). This finding is further supported by the factor that after the feature selection, when downsized to 100-112, the reduced full feature set has significantly improved in predicting adequacy and fluency than the 365 full feature set, and performs extremely well under SVM and RVM learning to estimate fluency in comparison to experiment 5, with significantly improved prediction accuracy (4.30→4.07 and 5.37→4.41 for adequacy, and 6.69→4.69 and 8.03→4.05 for fluency, with better results under RVM). However, support vector machine learned models with our feature set are comparatively complex (with significantly more numbers of support vectors as in experiment 7) and are likely to be less generalizable. Relevance vector machine models with our full feature set perform better than they are with the QuEst baseline feature set. From the above table, relevance vector machine does dramatically reduce the model complexity at the expense of affordable prediction accuracy loss in comparison to SVM. However, it is particularly noteworthy that word embedding features alone (200) can be high quality features for learning. It has achieved highest performance among all the experiments, of which only the fluency prediction in experiment 7 can be comparable to the embeddings as features.

The feature selection using Simulated Annealing returned 52 features as predicators for adequacy and 44 predicators for fluency from the 165 feature set that are largely features corresponding to grammatical and syntactical forms of the source and target texts. As far as we can see, these prominent features enjoy such special positions in the bilingual texts that they deserve the particular attention from trainee translators. For example, clause modifier, adverb modifier, coordination, parataxis, determiners, coordinating conjunction, and adposition are at marked slots of a sentence (F23-78). Take the sentence "One cannot deny that insects are a nuisance when their bites become sore, and a threat when they transmit disease, but, viewed dispassionately, even noxious insects are beautiful." as an example, it contains two coordinating conjunction (and, but), four adverb modifiers (when, even, dispassionately), one adverbial clause modifier (become and when) and two copulas.

A closer look at the first few translations suggests that trainee translators took divergent strategies to process and approach the text. Some neglect the coordinating relations, some under-translate the part in question and others express the internal logic connection of these elements in a rather clumsy way. As such, it may give raters an impression that the translations have space to improve. As we can see, trainee translators tend to neglect the coordinating conjunction of "nuisance" and "threat" and misinterpret the adverbial modifier "and even" to signify a logical contradiction. These explicit or implicit flow of meanings might be difficult points for human translators to grasp and influence their output quality. Another possible explanation is that the partial coincidence of target text features with source text features, like adjectives, noun or nouns phrases, verbs or verb phrases, conjunctions and even the particularly forma feature like auxiliary passives indicate that trainee translators, due to the lack of translation proficiency, are easily restricted by the source text forms and formal adherence become consciously or unconsciously the convenient translation strategy; Language modelling features, like out-of-vocabulary words, target text sentence LM log probability score and target text sentence part-of-speech log probability score, have suggested that translation quality, to some extent, depends upon how much of the translation is contained in the language model, and that adequacy and fluency sometimes are two such inseparable aspects that adequate translations often read smooth and native-like translations are more likely to be adequate.

The feature selection has yielded 100 features for adequacy and 112 features for fluency, and these selected features have significantly improved accuracy for both adequacy and fluency (with reduced cross validation errors) over the 365 full feature set. The selected features from the full feature set consist of mainly of bilingual ratio features, such as source text and target text indirect object ratio, adjective phrase ratio, adverbial phrase ratio, noun phrase ratio, semantic role label ratio, to just name a few, and overwhelmingly word embeddings (56) for adequacy and part-of-speech tags, dependency relations, syntactic trees and ratios of such tags between the source text and target text, in addition to the 57 word embeddings. This does make sense as word embeddings capture much semantic and synaptic information, they are useful information carrying the semantic and stylistic values across languages. In the meantime, though some part-of-speech tags are contributing to the predicting model, the majority of selected features seem to be larger units above word level, denoting that translation, while moving from the source text to the target text, is very likely to have some invariant relations at the a macro-level, for instance a direct object relation is of high possibility being translated as an direct object, and clausal subjects are often replaced with clausal subjects, which explains why such features from the source texts and target text are often selected coincidently. However, it shall be noted that we also compared the selected features obtained with different numbers of iterations (here we set 50, 100 iterations for two other round selection) while running feature annealing. The results show that each time only approximately 30 percent of the selected features will be retained regardless of the number of iterations set. This suggests a larger percentage of new features will be introduced when each time you run feature annealing to get a combination of new feature set.

4.3 Testing on Machine Translations

In order to validate our model, we test our model obtained in Experiment 7 on a new data set of Google Translate. We have the six ST texts translated by Google Translate and then extract the same feature set from the six machine translations as a new test data set. We compared the predicted adequacy and fluency scores under SVM and RVM for these six translations with 27 randomly sampled trainee translations from the original 277 training data. Table 6 shows the result (MT= Google translations, HT= student translations, Ad_S and Fl_S refers to adequacy and fluency scores with SVM learning, so does Ad_R and Fl_R meaning scores with RVM learning; sd= standard deviation, Sig. refers to significance level at 0.05.).

Data	Prd.	Mean	sd	Sig.
MT	Ad_S	38.04	2.00	$t_{(15.164)}$ = -0.89357
HT	Ad_S	39.04	3.94	p = 0.38
MT	Ad_R	27.51	5.77	$t_{(7.925)}$ = -4.0972
HT	Ad_R	38.37	6.33	p = 0.00
MT	Fl_S	27.20	1.94	$T_{(21.563)}$= 0.34583
HT	Fl_S	26.76	5.01	p = 0.73
MT	Fl_S	19.88	6.31	$T_{(6.6667)}$ = -2.452
HT	Fl_S	26.67	5.31	p = 0.04

Table 6 Model 7 Prediction Difference on MT & HT

Welch Two Sample t-tests demonstrate that RVM learned models are capable of differentiate machine translations and student translations, which are hypothetically deemed better than machine translations on the whole. This finding confirms with the result in Section 4.2 where in experiment 7 RVM models demonstrate improved simplicity in terms of fewer support vectors and equal and even higher accuracy in terms of cross validation error.

To illustrate the point about greater variability of human translation errors compared to MT we performed a pilot experiment using an error-annotation scheme of MeLLANGE [5] project to manually annotate the six machine translations and the 27student translations. While MT output typically contained errors related to content transfer, source language intrusion, and basic language errors, in contrast, human translation errors were more diverse, where over-translation, omission,

[5] http://mellange.eila.jussieu.fr/

word choice, language distortion, terminology errors and hygiene problems were very prominent.

In the meantime, we also examined the distribution of the predicted content and fluency scores. The results show that the predicted adequacy (mean=37.92, inter-quartile range =6.55 range=19.20) and the predicted fluency (mean =26.68, interquartile range=5.86, range=20.56) under RVM are very close to the interquartile range and range of the human scores. This suggests our model can correctly assign scores for most of the student translations.

5. Conclusion and Further Work

In this study, we investigated a feature set and a number of machine learning methods (SVM, RVM, feature selection and label propagation) to assess quality of human translations. We proposed our own feature set, which consists of four major components: monolingual features, bilingual features, language modelling features and bilingual embeddings. Monolingual features are designed to capture ST difficulty and TT fluency, language modelling features to assess TT fluency, bilingual features and word embeddings to assess the accuracy of transferring from ST to TT at different semantic and syntactic levels. We also compared our feature set to the QuEst set, which is commonly used in Quality Estimation for MT.

We also investigated feature selection methods on the original feature set to successively downsize it to a reasonably manageable one-third of the total number of features and have significantly boosted predicting adequacy and fluency respectively. Feature selection is of critical importance to build simpler yet almost equally effective model with large feature set and small sample size in our case.

Overall, through our experiments, we found the chosen machine learning algorithms perform better in predicting the fluency of human translations, and bilingual word embedding features, due to its advantage of capturing semantic and syntactic information, performs astonishingly well for adequacy and fluency prediction, with our feature set. The final model with our feature set developed at experiment 7 is capable of detecting difference between good and bad translations in most cases, as is discussed in Section 4.3.

Further work is need to predict the content quality of human translations in a more stable manner. Therefore, we will investigate more semantic-related features that approximate sentential and inter-sentential structures for transmission of meanings between the source and the target languages at different levels. Even though the performance of the bilingual word embedding features in our feature set together seems to be undermined by other features, we believe this feature set improves the interpretability of black-box nature of word embeddings and outperforms the formal-correspondence oriented QuEst feature set, with its easy accessibility and better explaining power to translation actions, is worth tinkering. From a practical point of view, we shall work out ways of better incorporating bilingual embeddings into the existing framework.

6. Acknowledgements

The first author is sponsored by the China Scholarship Council- University of Leeds Scholarship (Support No. [2013]3009) during this research.

7. Bibliographical References

Avramidis, E. (2012). Quality Estimation for Machine Translation Output Using Linguistic Analysis and Decoding Features. Proceedings of the Seventh Workshop on Statistical Machine Translation (pp. 84-94). Montreal, Canada: Association for Computational Linguistics.

Babych, B., & Hartley, A. (2008). Sensitivity of automated MT evaluation metrics on higher quality MT output: BLEU vs task-based evaluation methods. LREC08.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., & Specia, L. (2012). Findings of the 2012 Workshop on Statistical Machine Translation. Proceedings of the Seventh Workshop on Statistical Machine Translation, (pp. 10-51). Montreal, Canada.

Dhillon, P., Foster, D. P., & Ungar, L. H. (2011). Multi-view learning of word embeddings via cca. In Advances in Neural Information Processing Systems (pp. 199-207).

Dodigovic, M. (2005). Artificial intelligence in second language learning: Raising error awareness. Multilingual Matters.

Eisele, A., & Chen, Y. (2010, May). MultiUN: A Multilingual Corpus from United Nation Documents. In LREC.

Felice, M., & Specia, L. (2012). Linguistic Features for Quality Estimation. Proceedings fo the Seventh Workshop on Statistical Machine Translation (pp. 96-103). Montreal, Canada: Association for Computational Linguistics.

Fox, H. (2002). Phrasel Cohesion and Statistical Machine Translation. Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10 (pp. 304-311). Association for Computational Linguisitcs.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine learning, 46(1-3), pp. 389-422.

Blunsom, P., & Hermann, K. M. (2014). Multilingual Models for Compositional Distributional Semantics.

House, J. (2014). Translation Quality Assessment: Past and Present. Routledge.

Jiang, J. (2013). An automatic approach to evaluating the linguistic quality of English-Chinese translations. Modern Foreign Languages, 36(1), pp. 85-110.

Kirkpatrick, S. (1984). Optimization by simulated annealing: Quantitative studies. Journal of statistical physics, 34(5-6), pp. 975-986.

Ma, W.-y., & McKeown, K. (2013). Using a Supertagged

Dependency Language Model to Select a Good Translation in System Combination. HLT-NAACL, (pp. 433--438).

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014, June). he Stanford CoreNLP Natural Language Processing Toolkit. Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations (pp. 55-60). Baltimore, Maryland: Association for Computational Linguistics.

Owczarzak, K., van Genabith, J., & Way, A. (2007). Dependency-based Automatic Evaluation for Machine Translation. Proceedings of the NAACL-HLT 2007/AMTA Workshop on Syntax and Structure in Statistical Translation (pp. 80-87). Rochester, New York: Association for Computational Linguistics.

Padó, S., Cer, D., Galley, M., Jurafsky, D., & Manning, C. D. (2009). Measuring machine translation quality as semantic equivalence: A metric based on entailment features. Machine Translation, 23(2-3).

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th annual meeting on association for computational linguistics (pp. 311-318). Association for Computational Linguistics.

PetrovSlav, DasDipanjan, & McDonaldRyan. (04/2011). A Universal Part-of-Speech Tagset. ARXIV.

Rios, M., & Sharoff, S. (2015). Large Scale Translation Quality Estimation. Proceedings of the 1st Deep Machine Translation Workshop, (pp. 81-96). Prague, Czech Republic.

Shah, K., Avramidis, E., Biçicic, E., & Specia, L. (2013). QuEst–design, implementation and extensions of a framework for machine translation quality estimation. The Prague Bulletin of Mathematical Linguistics, 100, pp. 19-30.

Tian, L., Wong, D. F., Chao, L. S., Quaresma, P., Oliveira, F., & Yi, L. (2014). UM-Corpus: A Large English-Chinese Parallel Corpus for Statistical Machine Translation. In LREC (pp. 1837-1842).

Tian, Y. (2008, Feb.). On-line Automated Assessment of Translation. Chinese Science & Technology Translation Journal, 21(1).

Wen, Q. & Wang, J. (2008). Parallel Corpus of Chinese EFL Learners. Beijing : Foreign Language Teaching and Research Press.

Xiong, D., Zhang, M., & Li, H. (2010). Error detection for statistical machine translation using linguistic features. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 604-611). Association for Computational Linguistics.

Zhu, X., & Ghahramani, z. (2002). Learning from labeled and unlabelled data with label propagation. Carnegie Mellon University.