Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)

Assessing Translation Ability through Vocabulary Ability Assessment

Yo Ehara
AIST*
y-ehara@aist.go.jp

Yukino Baba
Kyoto University
baba@i.kyoto-u.ac.jp

Masao Utiyama, Eiichiro Sumita
NICT
{mutiyama,eiichiro.sumita}@nict.go.jp

Abstract

Translation ability is known as one of the most difficult language abilities to measure. A typical method of measuring translation ability involves asking translators to translate sentences and to request professional evaluators to grade the translations. It imposes a heavy burden on both translators and evaluators. In this paper, we propose a practical method for assessing translation ability. Our key idea is to incorporate translators' vocabulary knowledge for translation ability assessment. Our method involves just asking translators to tell if they know given words. Using this vocabulary information, we build a probabilistic model to estimate the translators' vocabulary and translation abilities simultaneously. We evaluated our method in a realistic crowdsourcing translation setting in which there is a great need to measure translators' translation ability to select good translators. The results of our experiments show that the proposed method accurately estimates translation ability and selects translators who have sufficient skills in translating a given sentence. We also found that our method significantly reduces the cost of crowdsourcing translation.

1 Introduction

Translation ability is known as one of the most difficult language abilities to measure [Stansfield *et al.*, 1992; Stubbe, 2014]. Such difficulty is caused by a variety of language proficiencies that comprise translation ability, for example, reading skills in the source language and writing skills in the target language. A typical method of measuring translation ability involves translation tests, in which a translator is asked to translate given sentences and professional evaluators are asked to grade the translation results. This method precisely assesses translation ability; however, hiring the professional evaluators is expensive and imposes a heavy burden on the translators.

*This work was performed for the National Institute of Information and Communications Technology (NICT) when the first author was employed by NICT. He currently works for the National Institute of Advanced Industrial Science and Technology (AIST).

There is a growing need for easy-to-use methods for measuring translation ability due to the recent expansion of crowdsourcing platforms that allow the hiring of non-professional translators. Because there is no guarantee that all non-professional translators on a crowdsourcing platform have sufficient translation skills, requesters who are willing to post translation jobs need to assess the translators' ability in order to obtain the services of high-quality translations and make good use of their budget.

In contrast with translation ability, vocabulary ability is considered a convenient measure for assessing language skills [Nation and Beglar, 2007; Meara, 2010]. For example, we can evaluate learners' vocabulary ability by using a simple test with which learners are asked to select the correct meaning from multiple choices. Such vocabulary tests have already been applied to select highly skilled second language learners [Nation, 2006] and for efficient student class placement [Read and Chapelle, 2001].

In this paper, we propose a convenient method for measuring translation ability. Our key idea is to use the results of low-cost vocabulary tests for translation ability assessment so that we do not need to rely on professional evaluators. Our method first obtains information about translators by making them take a translation test. In this test, the translators are asked to translate several sentences as well as to tell if they know the meaning of each word in the source language sentences. The second question is designed as a self-report vocabulary test and we use the results of the vocabulary test, called *vocabulary information* of translators, to estimate translation ability. Additionally, we can consider a case in which we can afford to hire professional evaluators to evaluate a small number of translations obtained from the translation test. Our method can be incorporated with such groundtruth information to estimate translation ability more precisely.

We introduce a probabilistic model for estimating translation ability. Our model is built based on an assumption that a translator with high translation ability is likely to have high vocabulary ability, and a translator with high vocabulary ability is likely to know the meaning of difficult words. Given the vocabulary information of translators, word difficulties and vocabulary and translation abilities are estimated using the maximum a posteriori (MAP) inference. Word features are incorporated for estimating word difficulties, therefore, when

a new source language sentence is given, our model can estimate the probability of a translator providing an acceptable translation for the sentence. Because we apply the self-report vocabulary test, one may be concerned that over-confident or unreliable translators are likely to answer that they know the meaning of most of the words. Our model incorporates a parameter representing unreliability of a translator to address such cases.

We conducted two experiments indicating that the proposed method accurately estimates translation ability. The first experiment showed that the proposed method selected good translators in a realistic setting of crowdsourcing translation. The results will immediately lead to an application to reduce the cost of crowdsourcing translation. The second experiment showed that estimated vocabulary and translation ability correlate, which validates our method.

The contributions of this paper are as follows:

- This is the first study addressing easy translation ability measurement using vocabulary measurement.
- From investigating the obtained model, we argue that translation ability correlates with vocabulary ability.
- Our method significantly reduces the cost in selecting good translators in a realistic setting of crowdsourcing translation.

2 Problem setting

Let $\mathcal{S}$ be a given set of source language sentences for translation tests. Each source sentence $s \in \mathcal{S}$ is defined as an enumeration of words $s = \{w_{s1}, w_{s2}, \cdots\}$. Each word w in a source language is associated with an N-dimensional feature vector $\boldsymbol{w} \in \mathbb{R}^N$ and each sentence s is associated with an M-dimensional feature vector $\boldsymbol{s} \in \mathbb{R}^M$.

Let the set of translators be $\mathcal{K}$. Each translator $k \in \mathcal{K}$ takes a translation test and provides translations for a set of source language sentences $\mathcal{S}_k$. The quality of a translation given by translator k for source language sentence s is then graded by a professional evaluator. Let $z_{ks} = 1$ if the translation is judged as *acceptable* and $z_{ks} = 0$ otherwise. We call z_{ks} the *translation quality label*. We consider two settings: *semi-supervised* and *unsupervised*. In the semi-supervised setting, we assume that some test translations is graded by professional evaluators while none are evaluated in the unsupervised setting, that is, z_{ks} is not given for all $k \in \mathcal{K}$ and $s \in \mathcal{S}_k$.

When a translator k translates a test sentence, she is also asked to tell if she knows the meaning of each word in the sentence. If k answers that she knows the meaning of word w, then we set $y_{kw} = 1$; otherwise, $y_{kw} = 0$. We assume that y_{kw} is given for all $k \in \mathcal{K}, s \in \mathcal{S}_k$, and $w \in s$.

Given $\{\boldsymbol{s}\}_{s\in\mathcal{S}}$, $\{\boldsymbol{w}\}_{w\in s, s\in\mathcal{S}}$, $\{y_{kw}\}_{k\in\mathcal{K}, s\in\mathcal{S}_k, w\in s}$, and $\{z_{ks}\}_{k,s}$, our goal was to build a translation quality predictor $\Pr\left(z_{ks'} = 1 | \boldsymbol{s}', \{\boldsymbol{w}\}_{w\in s'}, \{y_{kw}\}_{w\in s'}\right)$, where $s' \notin \mathcal{S}$ is a new source language sentence. When translator k provides her vocabulary knowledge for every word $w \in s'$, the predictor predicts whether the quality of translation given by k for s' will be acceptable or not, even when k does not actually translate the sentence. Notations are listed in Table 1.

3 Vocabulary Knowledge-based Translation Quality Predictor (VKTQP)

3.1 Translation Ability Model

Our model called Vocabulary Knowledge-based Translation Quality Predictor (VKTQP) is composed of two probabilistic models: *translation ability model* and *vocabulary ability model*. For building the translation ability model, we assume that translation quality varies according to translation ability of a translator, and translation difficulty of a source language sentence. The Rasch model [Rasch, 1960; Baker and Kim, 2004] holds the key idea of prediction regarding estimating ability and difficulty. This is a simple and standard model for representing human ability and task difficulty in education and psychology. We model the translation ability of translator k as ψ_k and translation difficulty of source language sentence s as ν_s.

Using the Rasch model, the probability that translator k acceptably translates source language sentence s is given as

$$\Pr\left[z_{ks} = 1\right] = \sigma\left(\psi_k - \nu_s\right), \tag{1}$$

where σ denotes the logistic sigmoid function, i.e., for a real number a, $\sigma(a) \equiv \frac{1}{\exp(-a)+1}$. Being binary, z_{ks} is predicted to be 1 if translation ability ψ_k outperforms difficulty ν_s; otherwise, 0. Note that the Rasch model is a special case of logistic regression.

We further model sentence difficulty ν_s as the inner product of sentence feature vector $\boldsymbol{s}$ and weight parameter $\boldsymbol{\tau}$, that is, $\nu_s = \boldsymbol{\tau}^\top \boldsymbol{s}$. Incorporating sentence feature vectors allows our model to estimate the probability of a translator providing an acceptable translation for a new sentence. The translation ability model is illustrated on the left side of Figure 1.

3.2 Vocabulary Ability Model

We assume that the probability of a translator knowing the meaning of a word is dependent on the vocabulary ability of the translator and difficulty of the word. We model the vocabulary ability of translator k as θ_k and difficulty of word w as μ_w. Because we apply the self-report vocabulary test, one may be concerned that over-confident or unreliable translators are likely to answer that they know the meaning of most of the words. Thus, we introduce a parameter ϕ_k to model

Table 1: Notations of observed variables and model parameters

k	translator index
s	source language sentence index
w	source language word
$\mathcal{S}$	set of source language sentences
$\mathcal{S}_k$	set of source language sentences assigned to k
$\boldsymbol{s}$	feature vector of s
$\boldsymbol{w}$	feature vector of w
y_{kw}	$\in \{0, 1\}$. 1 if k knows w; otherwise, 0
z_{ks}	$\in \{0, 1\}$. 1 if k translates s at an acceptable quality; otherwise, 0
ψ_k	$\in \mathbb{R}$. k's translation ability.
θ_k	$\in \mathbb{R}$. k's vocabulary ability.
ϕ_k	$\in \mathbb{R}$. Unreliability of translator k.
$\boldsymbol{\tau}$	weight vector for difficulty in translating s
$\boldsymbol{\pi}$	weight vector for difficulty of w

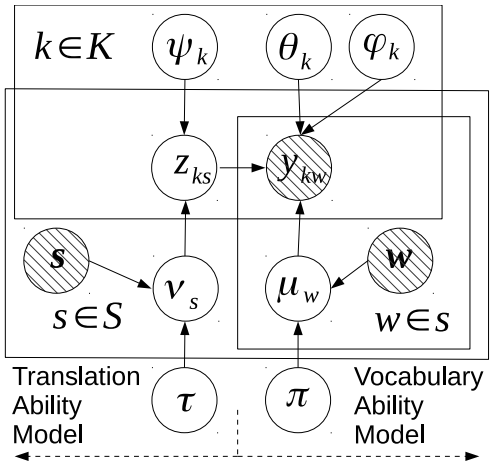

Figure 1: Graphical representation of VKTQP. Hatched circles denote observed variables. Left side represents translation ability model and right side shows vocabulary ability model.

the unreliability of translator k. We assume that, if translator k provides an acceptable translation for sentence s, the probability of k answering that she knows the meaning of word $w \in s$ depends on vocabulary ability θ_k. Otherwise, we model the probability by using ϕ_k. We apply the Rasch model and represent the probabilities, namely,

$$\Pr[y_{kw} = 1 \mid z_{ks} = 1] = \sigma(\theta_k - \mu_w) \quad (2)$$

$$\Pr[y_{kw} = 1 \mid z_{ks} = 0] = \sigma(\phi_k - \mu_w), \quad (3)$$

where $w \in s$.

We also incorporate the word feature vectors into the vocabulary ability model by modeling the difficulty of w as $\mu_w = \boldsymbol{\pi}^\top \boldsymbol{w}$, where $\boldsymbol{\pi}$ denotes a weight parameter. The vocabulary ability model is illustrated on the right side of Figure 1.

Overall, the probability described with the VKTQP model is written as

$$\prod_k \prod_{s \in \mathcal{S}_k} \prod_{w \in s} \Pr[y_{kw} \mid z_{ks}] \Pr[z_{ks}], \quad (4)$$

where $\Pr[y_{kw} = 1 \mid z_{ks}] = \sigma(\theta_k - \mu_w)^{z_{ks}} \sigma(\phi_k - \mu_w)^{1 - z_{ks}}$.

4 Parameter Estimation

4.1 Priors

We now define the priors on the parameters listed in Table 2. Since we have many parameters, priors are important not only for preventing overfitting but also for describing relations between parameters.

First, for preventing overfitting, we define the priors for $\boldsymbol{\pi}$, $\boldsymbol{\tau}$, θ_k, and ϕ_k. In Table 2, $\mathcal{N}$ denotes the probability density function of the Gaussian distribution. These priors are set to have 0 means so that they can prevent overfitting by regularizing parameters towards 0.

Second, we define two versions of the VKTQP models called VKTQP-I and VKTQP-D by changing the prior for translation ability ψ_k. These models have a direct effect on the prediction of translation quality (Table 2). The VKTQP-I model simply uses the Gaussian priors with a 0 mean for preventing overfitting for each parameter. Unlike VKTQP-I, the prior for ψ_k in VKTQP-D is set to θ_k. This prior enforces the assumption that translation ability will correlate with vocabulary ability by regularizing translation ability ψ_k towards vocabulary ability θ_k.

Table 2: Priors of VKTQP models. By changing priors of ψ, two models are constructed. λ and ξ are hyper-parameters.

$\boldsymbol{\pi}$		$\Pr[\boldsymbol{\pi}] \equiv \mathcal{N}(\mathbf{0}, \lambda^{-1}\boldsymbol{I})$
$\boldsymbol{\tau}$		$\Pr[\boldsymbol{\tau}] \equiv \mathcal{N}(\mathbf{0}, \xi^{-1}\boldsymbol{I})$
θ_k		$\Pr[\theta_k] \equiv \mathcal{N}(0, \lambda^{-1})$
ϕ_k		$\Pr[\phi_k] \equiv \mathcal{N}(0, \lambda^{-1})$
ψ	VKTQP-I	$\Pr[\psi_k] \equiv \mathcal{N}(0, \xi^{-1})$
	VKTQP-D	$\Pr[\psi_k] \equiv \mathcal{N}(\theta_k, \xi^{-1})$

4.2 Parameter Estimation of Supervised Setting

We use the MAP inference for estimating the model parameters. The posterior distribution is written as

$$\begin{aligned} &L(\{\theta_k\}_K, \{\phi_k\}_K, \{\psi_k\}_K, \boldsymbol{\pi}, \boldsymbol{\tau}) \\ &= \text{Priors}[\{\theta_k\}_K, \{\phi_k\}_K, \{\psi_k\}_K, \boldsymbol{\pi}, \boldsymbol{\tau}] \\ &\times \prod_k \prod_{s \in \mathcal{S}_k} \prod_{w \in s} \Pr[y_{kw} \mid z_{ks}] \Pr[z_{ks}], \end{aligned} \quad (5)$$

where the function "Priors" returns the products of the priors of each parameter defined in the previous section.

Instead of maximizing (5), we minimize its negative log of (5) for the MAP inference. Although we only consider the unsupervised and semi-supervised settings, it would be worth mentioning that when we observe z_{ks} for all k and $s \in \mathcal{S}_k$ (i.e., in the supervised setting), the negative log of the posterior is convex with respect to all parameters. This means that we can obtain the optimal parameters. This convexity is directly derived from that of the negative log of the posterior function of the L2-normalized logistic regression.

4.3 Parameter Estimation in Unsupervised and Semi-supervised Settings

In the unsupervised and semi-supervised settings, the set of translation qualities $\{z_{ks}\}$ contain unobserved variables. Thus, we use the expectation-maximization (EM) algorithm [Dempster and Rubin, 1977] for the MAP inference of the parameters. Once the unsupervised setting is derived, the semi-supervised setting can be easily derived by fixing the observed part of $\{z_{ks}\}$ to the observed values.

For simplifying the equations, we introduce the following notations. α_{kw} denotes the probability that k knows w given that z_{ks} is acceptable, i.e., $\alpha_{kw} \equiv \Pr[y_{kw} = 1 \mid z_{ks} = 1] = \sigma(\theta_k - \mu_w)$; β_{kw} denotes the probability that k knows w given that z_{ks} is not acceptable, i.e., $\beta_{kw} \equiv \Pr[y_{kw} = 1 \mid z_{ks} = 0] = \sigma(\phi_k - \mu_w)$; and γ_{ks} denotes the probability of obtaining an acceptable quality when k translates s, i.e., $\gamma_{ks} \equiv \Pr[z_{ks} = 1] = \sigma(\psi_k - \nu_s)$.

In the E-step, we evaluate the following formula to update z_{ks}.

$$\Pr\left[z_{ks}=1 \mid \{y_{kw}\}_{w\in s}\right] \propto \gamma_{ks}\Big(\prod_{w\in s}\alpha_{kw}^{y_{kw}}\left(1-\alpha_{kw}\right)^{(1-y_{kw})}\Big)$$

$$\Pr\left[z_{ks}=0 \mid \{y_{kw}\}_{w\in s}\right] \propto (1-\gamma_{ks})\Big(\prod_{w\in s}\beta_{kw}^{y_{kw}}\left(1-\beta_{kw}\right)^{(1-y_{kw})}\Big)$$

After the E-step, we update z_{ks} with new z'_{ks}, which is defined as $z'_{ks} \equiv \Pr\left[z_{ks}=1 \mid \{y_{kw}\}_{w\in s}\right]$.

In the M-step, we maximize the Q function, which has the same form as the likelihood function L in (5), except that z_{ks} is replaced with the updated value z'_{ks}. Once z'_{ks} is replaced and fixed, the negative log of the Q function is convex with respect to parameters, as is the case with the negative log likelihood. Thus, the Q function can be globally optimized in the M-step.

5 Evaluation

5.1 Dataset

We targeted English-to-Japanese translation for our experiments. We used the Japanese-English Bilingual Corpus of Wikipedia's Kyoto Articles[1]. We randomly selected 104 English sentences that have more than ten words and their corresponding translations given in Japanese.

For collecting a dataset, we used Lancers [2] crowdsourcing service, one of the largest crowdsourcing services in Japan. Most of the translators working in this service were assumed to be native Japanese speakers.

We hired 55 translators and obtained 1,498 English-to-Japanese translations. The translation cost was 10 JPY (approximately 0.1 USD) for each sentence. One sentence was translated by 14.4 translators on average. One translator translated 27.2 sentences on average.

Before translators engaged in the English-to-Japanese translation tasks, we asked the translators to report the words unfamiliar to them by clicking on them. Then, the translators started their translation tasks. The use of dictionaries was allowed (but not required) during the tasks.

The translation quality used for the evaluation was manually judged by two annotators, who were native speakers of the target language (Japanese) and fluent in the source language (English). The translation quality was judged using a 5-point scale, where 5 is acceptable and the other 4 points are unacceptable translation mistakes. The kappa coefficient[3] on the acceptability of the translations between the two annotators was 0.619, which is a "significant agreement" [Landis and Koch, 1977; Mihalcea and Chklovski, 2004].

[1]http://alaginrc.nict.go.jp/WikiCorpus/index_E.html

[2]http://www.lancers.jp

[3]We used the Python NLTK library for calculating the kappa coefficient. http://nltk.org/

Table 3: List of sentence features

N-gram Perplexity	Perplexity using 3-gram language model. Model was created using standard procedure for creating English language model in Moses toolkit using News-Commentary Corpus [4].
OOV words	Number of out-of-vocabulary (OOV) words in a sentence. We regard words that do not appear in trained language model above as OOV words.
# of words	Number of words in sentence.
# of commas	Number of commas in sentence.
Rates	Per-word rates were also put into features: namely, perpelxity/# of words, OOV words/ # of words, # of commas/ # of words

Table 4: List of word features

Google-1 gram [5]	Negative log of 1-gram probability of words in the Web 1T 5-gram Corpus
COCA [6]	Negative log of 1-gram probability of words in Corpus of Contemporary American English (COCA).
Brown corpus	Negative log of 1-gram probability of words in Brown Corpus.
SVL 12000 [7]	Manually labeled word difficulty index that ranges from 1, easiest, to 12, most difficult. This difficulty index is designed to measure difficulty for Japanese-English learners.

5.2 Features

Throughout the experiments, all our VKTQP models used the same sentence features designed to allow us to easily apply our method to other source languages. Table 3 lists the sentence features.

The word features use three corpora and one difficulty index (Table 4). These features were reported to be helpful for accurately predicting whether a learner knows a word [Ehara *et al.*, 2012; 2013; 2014].

5.3 Experiments

Table 5 lists the models that we compared in our experiments. VKTQP-I and VKTQP-D are our models.

We used five-fold nested cross validation throughout the experiments. All the models have hyper-parameters. We conducted grid-search over the validation sets for tuning the hyper-parameters. The hyper-parameters of all the models were chosen from $10^{-3.0}$, $10^{-2.4}$, $10^{-1.8}$, $10^{-1.2}$, $10^{-0.6}$, $10^{0.0}$, $10^{0.6}$, $10^{1.2}$, $10^{1.8}$, $10^{2.4}$, and $10^{3.0}$. After tuning the hyper-parameters over the validation set, we measured the performance on the test set, which is disjoint from the validation and training sets.

5.4 Cost Reduction Rate: Evaluation Measure for Assigning Tasks

As a realistic evaluation measure in crowdsourcing translation, we define a measure called the cost reduction rate (CRR). Cost indicates the number of translation requests of

[4]http://www.statmt.org/moses/?n=moses.baseline

[5]https://catalog.ldc.upenn.edu/LDC2006T13

[6]http://corpus.byu.edu/coca/

[7]http://www.alc.co.jp/vocgram/article/svl/

Table 5: List of models used in experiment. VKTQP-I, and VKTQP-D are our models. Others are baseline.

SVM	Support vector machine (SVM) with only sentence features.
LR	Logistic regression (LR) with only sentence features. This case can be regarded as that in which Rasch model is used for predicting translation quality.
SVM+AVGWF	SVM using both sentence features and averaged vocabulary feature vectors. We averaged word feature vectors of all known words and those of unfamiliar words in each sentence.
LR+AVGWF	LR using both sentence features and averaged word feature vectors.
VKTQP-I	Our model with independent prior
VKTQP-D	Our model with prior enforcing translation ability to be close to VKTQP

translators to translate a new sentence. The CRR shows a decrease in the number of requests by the predictor's assignments compared to random assignments.

Suppose we have a new sentence to translate, a set of available translators, and a predictor. First, all the available translators are candidate translators. The predictor ranks the translators by its predicted translation quality. We select the best translator from the predictor's ranking and request translation of the sentence from him/her. Then, we evaluate the translator's translation quality. If the quality is not acceptable, we discount the translator from the group of candidates and request the next translator to translate the same sentence. We repeat these steps until we encounter a translator who returns an acceptable translation. The number of translators until then is defined as #TR.

The CRR is defined in (6). Larger CRR values mean larger reduction or better performance. The numerator of (6) is the number of requests by the predictor until an acceptable translation is obtained. The denominator of (6) is the expected number of requests by random predictors. In other words, the denominator is the average number of requests until an acceptable translation is obtained when choosing translators randomly.

$$CRR \equiv 1 - \frac{\text{predictor's \#TR}}{E[\text{random predictors' \#TR}]}. \quad (6)$$

For example, suppose only 1 in 3 translators can translate a given source sentence acceptably. If we randomly take translators from the bag of 3 without replacement, $1/3 * (1 + 2 + 3) = 2$ translators are necessary to obtain one good translation on average. If we encounter a good translator first, the CRR is $1 - 1/2 = 1/2$, which halves the cost. If second, the CRR is $1 - 2/2 = 0$, which provides no improvement. If third, the CRR is $1 - 3/2 = -1/2$, which increases the cost.

If no translator can translate the sentence acceptably, we define CRR as 0. If all translators can translate the sentence acceptably, CRR also becomes 0. This means that no improvement can be made by any predictor.

5.5 Evaluation based on Cost Reduction Rate

Table 6 shows the CRR values of each model. Each column shows the size of the training data. The column with 0 training data shows the experiments in an unsupervised setting, where no translation quality label was provided. We used 5-fold nested cross-validation for this evaluation as well.

Table 6: Cost reduction rate of each model against training size. In each column, bold value denotes best model and underlined value denotes second best. Asterisks denote that best model significantly outperformed second best model. (**:$p < .01$, *:$p < .10$, Wilcoxon test was used).

Training size	0 (0%)	10 (1%)	55 (6%)	112 (13%)	896 (100%)
SVM	-	0.175	0.297	0.328	0.326
LR	-	0.180	0.294	0.331	0.329
SVM+AVGWF	-	0.244	0.321	**0.341****	<u>0.347</u>
LR+AVGWF	-	<u>0.269</u>	<u>0.322</u>	<u>0.340</u>	**0.355****
VKTQP-I	<u>0.092</u>	0.264	0.301	0.339	0.334
VKTQP-D	**0.327****	**0.340****	**0.325****	0.336	0.337

In Table 6, in an unsupervised setting where the training size is 0, we can easily see that VKTQP-D reduced cost by 32.7%. This is our main result. The reason that translation ability was difficult to estimate in previous studies lies in the high cost for obtaining translation quality labels. The reduction in cost shows that VKTQP-D estimated translation ability accurately through vocabulary ability using no translation quality labels. The estimation was so accurate that it would actually reduce the cost of crowdsourcing translation by selecting good translators.

The next notable observation is that VKTQP-D consistently outperformed VKTQP-I. As explained earlier, VKTQP-D differs from VKTQP-I in that with VKTQP-D, it is assumed that translation ability is close to vocabulary ability. Thus, this observation demonstrates that this assumption is true and vocabulary tests are helpful for predicting a translator's translation ability. This difference between VKTQP-Ds and VKTQP-Is was statistically significant where the training size was 0, 10, and 55.

The SVM+AVGWF constantly outperformed SVM and LR+AVGWF constantly outperformed LR. The former, i.e., those with "AvgWF" in their names, use vocabulary features while the latter do not. This result shows that the use of vocabulary features improves predictive performance even if the same algorithms, i.e., SVM and logistic regression, are used.

5.6 Analysis of Abilities

The previous section supports the assumption that making translation ability ψ close to vocabulary ability θ achieves better performance. This section further verifies this assumption by directly observing the translation ability ψ and vocabulary ability θ.

The values of translation ability ψ are plotted against those of vocabulary ability θ. In each sub-figure of Figure 2, each dot represents each translator. The dots were regressed by linear regression. The regression results are shown in the caption of each sub-figure.

In Figure 2, the upper figures, namely a) and b), show the results when the ability parameters were estimated without any translation quality labels, whereas c) and d) show those when the ability parameters were estimated using all 896 translation quality labels.

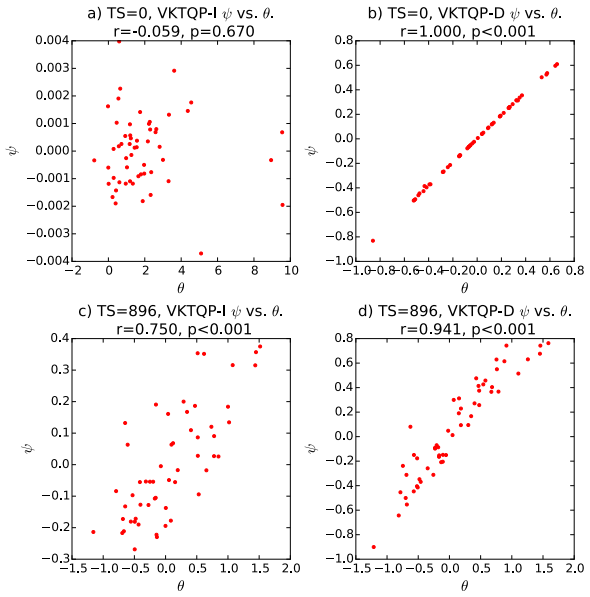

Figure 2: Plots of ability parameters. TS denotes training size. Each dot represents each of 55 translators in each subplot. r denotes correlation coefficient and p denotes p-value under null hypothesis in which $r = 0$.

When the training size (TS) was 0 in Figure 2, we can clearly see that, in a), there is no correlation between translation ability ψ and vocabulary ability θ. In contrast, in b), there is a clear correlation between translation ability ψ and vocabulary ability θ. This is because VKTQP-D used in b) is designed to make translation ability close to vocabulary ability, whereas VKTQP-I is not designed to do so. Note that VKTQP-D greatly outperformed VKTQP-I in this unsupervised setting, as stated in the previous section.

An interesting observation is shown in the lower figures of Figure 2, when the TS was 896. Although VKTQP-I is not designed to make translation ability ψ and vocabulary ability θ close, these two abilities estimated with VKTQP-I in c) clearly correlated. In fact, this correlation is statistically significant. These results verify the assumption that translation ability and vocabulary ability naturally correlate. Since this is directly assumed with VKTQP-D and this assumption is actually true, VKTQP-D outperforms VKTQP-I.

When the TS was 896, the dots of VKTQP-D were more scattered in d) compared to those in b) while the dots in d) still significantly correlated. The reason of this is presumably due to the noise in the translation quality labels.

6 Related Work

The studies [Malakoff, 1992] and [Stansfield *et al.*, 1992] are seminal in translation ability measurement. They involved measuring translation ability in realistic job-related settings in Spanish-to-English translation. While they focused on professional translations or translation by bilingual people, they did not address non-professional translations by second-language learners.

Stubbe [2014] recently addressed the relationship between translation ability and vocabulary ability at the word level. However, he did not address sentence-level translation nor propose a method for measuring translation ability by using vocabulary ability. Gao *et al.* [2015] recently addressed reducing costs in crowdsourcing translations. However, they measured translation ability by actually letting translators translate some sentences for a trial. Thus, they did not mention a method for reducing the cost of measuring translation ability. Moreover, they did not address vocabulary measurement since it was not their focus.

Previous related studies focused on estimating the quality of *given* translations, while we aimed to predict translation quality *without* translations. Quality Estimation (QE) is an example of these studies, which aims to estimate the quality of given translations. Quality estimation is typically formulated as a supervised regression problem [de Souza *et al.*, 2014]. It also involves crowdsourcing for collecting supervision [Zaidan and Callison-Burch, 2009; 2010] and has been extensively studied, e.g., [Callison-Burch *et al.*, 2012; Bojar *et al.*, 2013]. Quality estimation mainly targets machine-translated texts. In contrast, application of QE to human translation is very recent [Specia and Shah, 2014]. Although in that study they used the verb "predict", their approach, unlike ours, requires written translations for estimating quality. For clarity, we use "predict" only when translations are not given and "estimate" only when they are.

Regarding our use of the self-reporting approach for measuring vocabulary, many educational experts support self-report vocabulary measurement, as surveyed by Nation [2006]. Typical multiple-choice measurement, in which learners are asked to choose the correct meaning of a word from a set of multiple choices, depends heavily on the creation of "non-correct" choices. Therefore, those who support self-report vocabulary measurement claim that it is more reliable than multiple-choice measurement [Meara, 2010].

7 Conclusion

Translation ability is difficult and costly to measure compared to vocabulary ability, though there is a great need to measure it. This paper is the first to propose a method for measuring translation ability through the measurement of vocabulary ability and requires little or no translation.

We quantitatively evaluated our method in a realistic setting where the measured translation ability was used to select good translators for crowdsourcing translation. Our method exhibited significantly high accuracy in selecting translators who produce acceptable translations.

Future work includes support for other types of information, such as editing information, to measure translators' translation ability.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 26730115 and 15K16059.

References

[Baker and Kim, 2004] Frank B. Baker and Seock-Ho Kim. *Item response theory: parameter estimation techniques*. CRC Press, second edition, 2004.

[Bojar *et al.*, 2013] Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Findings of the 2013 workshop on statistical machine translation. In *Proceedings of the 8th Workshop on Statistical Machine Translation (WMT)*, pages 1–44, 2013.

[Callison-Burch *et al.*, 2012] Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the 7th Workshop on Statistical Machine Translation (WMT)*, pages 10–51, 2012.

[de Souza *et al.*, 2014] Jose G. C. de Souza, Marco Turchi, and Matteo Negri. Machine translation quality estimation across domains. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, pages 409–420, 2014.

[Dempster and Rubin, 1977] Nan M. Laird Dempster, Arthur P. and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[Ehara *et al.*, 2012] Yo Ehara, Issei Sato, Hidekazu Oiwa, and Hiroshi Nakagawa. Mining words in the minds of second language learners: learner-specific word difficulty. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, 2012.

[Ehara *et al.*, 2013] Yo Ehara, Nobuyuki Shimizu, Takashi Ninomiya, and Hiroshi Nakagawa. Personalized reading support for second-language web documents. *ACM Transactions on Intelligent Systems and Technology*, 4(2), 2013.

[Ehara *et al.*, 2014] Yo Ehara, Yusuke Miyao, Hidekazu Oiwa, Issei Sato, and Hiroshi Nakagawa. Formalizing word sampling for vocabulary prediction as graph-based active learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1374–1384, 2014.

[Gao *et al.*, 2015] Mingkun Gao, Wei Xu, and Chris Callison-Burch. Cost optimization in crowdsourcing translation: Low cost translations made even cheaper. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 705–713, 2015.

[Landis and Koch, 1977] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174, 1977.

[Malakoff, 1992] Marguerite E. Malakoff. Translation ability: A natural bilingual and metalinguistic skill. *Advances in Psychology*, 1992.

[Meara, 2010] Paul M. Meara. *EFL Vocabulary Tests (Second Edition)*. Swansea: Centre for Applied Language Studies, 2010.

[Mihalcea and Chklovski, 2004] Rada Mihalcea and Timothy Chklovski. Building sense tagged corpora with volunteer contributions over the web. *Recent Advances in Natural Language Processing III: Selected Papers from RANLP 2003*, 260:357, 2004.

[Nation and Beglar, 2007] I.S.P. Nation and David Beglar. A vocabulary size test. *The Language Teacher*, 31(7):9–13, 2007.

[Nation, 2006] I. S. P. Nation. How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, 63(1):59–82, 2006.

[Rasch, 1960] Georg Rasch. *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen, Denmark: Denmarks Paedagogiske Institut, 1960.

[Read and Chapelle, 2001] John Read and Carol A. Chapelle. A framework for second language vocabulary assessment. *Language Testing*, 1:1–32, 2001.

[Specia and Shah, 2014] Lucia Specia and Kashir Shah. Predicting human translation quality. In *Proceedings of the eleventh biennial conference of the Association for Machine Translation in the Americas (AMTA)*, pages 288–300, 2014.

[Stansfield *et al.*, 1992] Charles W. Stansfield, Mary Lee Scott, and Dorry Mann Kenyon. The measurement of translation ability. *The Modern Language Journal*, 1992.

[Stubbe, 2014] Raymond Stubbe. Do japanese students overestimate or underestimate their knowledge of english loanwords more than non-loanwords on yes-no vocabulary tests? *Vocabulary Learning and Instruction*, 2014.

[Zaidan and Callison-Burch, 2009] Omar F Zaidan and Chris Callison-Burch. Feasibility of human-in-the-loop minimum error rate training. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 52–61, 2009.

[Zaidan and Callison-Burch, 2010] Omar F Zaidan and Chris Callison-Burch. Predicting human-targeted translation edit rate via untrained human annotators. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2010.