Cross-lingual Terminology Extraction for Translation Quality Estimation*

Yu Yuan[†] [*], Yuze Gao[‡], Yue Zhang[‡], Serge Sharoff[*]
[†] School of Languages & Cultures, Nanjing University of Information Science & Technology, 210044, China
[‡] Information Systems Technology and Design, Singapore University of Technology and Design, 487372
[*] Centre for Translation Studies, University of Leeds, LS2 9JT, United Kingdom
hittle.yuan@gmail.com, {yuze_gao, yue_zhang}@sutd.edu.sg s.sharoff@leeds.ac.uk

Abstract

We explore ways of identifying terms from monolingual texts and integrate them into investigating the contribution of terminology to translation quality.The researchers proposed a supervised learning method using common statistical measures for termhood and unithood as features to train classifiers for identifying terms in cross-domain and cross-language settings. On its basis, sequences of words from source texts (STs) and target texts (TTs) are aligned naively through a fuzzy matching mechanism for identifying the correctly translated term equivalents in student translations. Correlation analyses further show that normalized term occurrences in translations have weak linear relationship with translation quality in term of usefulness/transfer, terminology/style, idiomatic writing and target mechanics and near- and above-strong relationship with the overall translation quality. This method has demonstrated some reliability in automatically identifying terms in human translations. However, drawbacks in handling low frequency terms and term variations shall be dealt in the future.

Keywords: Bilingual terminology, translation quality, supervised learning, correlation analysis

1. Introduction

Terminology helps translators organize their domain knowledge, and provides them means (usually terms in various lexical units) to express subject knowledge adequately. Translation scholars and practitioners maintain that terminology correctness is associated with the quality of translation (and interpretation) (Hartley et al., 2004; Xu and Sharoff, 2014; Kim et al., 2015; Brunette, 2000; Karoubi, 2016).

The acknowledgement of the contribution of terminology to translation quality is also echoed by the translation industry and users (Secară, 2005; Lommel et al., 2014; Warburton, 2013). Accurately reproducing the content of the original and using appropriate terminology has become the official assessment criteria of some famous in-use translation-error-based evaluation schemes. For instance, the MeLLANGE project (Secară, 2005) defines more than six terminology errors[1], and the Multidimensional Quality Metrics lists terminology as one of the eight major dimensions, which is subdivided into three children issue types (term inconsistency, termbase[2], and terminology domain[3]) (Lommel et al., 2014). From a user's expectation perspective, appropriate terminological use is also viewed as one of the important quality parameters. For the purpose of marketing, companies will localize the manuals that accompany their products. Localization cannot be done at the expense of quality to endanger the customer satisfaction. Their dissatisfaction will lead to more potential damaging losses in revenue. Therefore, speed and quality is what localization services users are looking for (Warburton, 2013).They would expect that all the terms are translated correctly and consistently, and translators will not invent terms randomly wherever source language (SL) terms cannot find an equivalent in target language (TL) without scientific analysis and sufficient documentation. For both sides, adherence to specified terminology is considered a central concern in translation for the delivery of quality-assured translations.

It is clear that finding an equivalent for terms in a translation impacts the overall quality of translation. When assessing a translation, evaluators should consider how well a translator achieves in successfully rendering those terms in the target language. However, this element of translation has not drawn enough attention from researchers in machine translation quality estimation, and in human translation quality assessment, the whole evaluation of the translation of terminology is carried out by human evaluators manually and subjectively, with or without references. Manual compilation of bilingual term lists for each translation evaluation task is an expensive and laborious effort, hence the rarity of an up-to-date, specialized and relatively comprehensive term database for translation quality estimation purpose.

The main contributions of our work include: language and model adaptation by training term classifiers using a corpus in the bio-medical domain and applying the optimal classifiers to cross-domain and cross-language texts; investigating the contribution of terminology to translation quality with empirical evidence; a working pipeline for terminology-focused quality evaluation to extract and exploit terminology information from raw source texts (STs) and target texts (TTs).

2. Related Work

Different from monolingual term extraction, bilingual term extraction (BTE) faces the additional problem of finding translation equivalents in parallel or comparable texts.

This work is done when the first author works as a research fellow at SUTD.

[1]The main terminological errors are incorrect terminology, false cognate, term translated by non-term, inconsistent with glossary, inconsistent within target text (TT), inappropriate collocation, and user-defined errors.

[2]a term is translated is translated with a term nonconforming to the specification.

[3]a term is translated with a term from a different domain.

There are roughly three approaches to bilingual term extraction, depending on what resources are used:

- **Parallel-corpus Based** Various strategies (Gómez Guinovart and Simoes, 2009; Macken et al., 2013) have been advanced for extracting lexical equivalence from parallel corpora. The main fallacy of methods in this approach is that they rely on the morphosynactic analyser of the term extractor that does not recognize all candidate terms and those chunk-based methods, having extended the alignment model with automatically extracted language pair specific rules. As a consequence, this method blurs the distinction between terms and non-terms.

- **Comparable-corpus Based** Bilingual corpora in specialized domains are actually scarce and it is expensive to build high quality parallel texts of specialized domains. A practical solution to this limitation is to make use of comparable corpora (Rocheteau and Daille, 2011; Xu et al., 2015; Hakami and Bollegala, 2017) that are available in large quantities. However, term extraction along this line is often limited to noun phrases (< 5 words) from monolingual comparable corpora. Thus, the recall of such an approach is not satisfactory under some circumstances. For other studies in this approach, ambiguity of term translations and identification of synonymous terms need to be further addressed.

- **Web-data Based** Web data mining is another means to collect terminology pairs (Erdmann et al., 2009; Gaizauskas et al., 2015). Despite the favourable findings from the evaluation process, one of the biggest limitations of the current approach is that the precision still warrants improvement in comparison to other methods that are parallel-corpus based.

To sum up, these systems and pipelines are designed for terminology management or dictionary compilation purpose rather than translation quality evaluation. They cannot readily serve our purpose of finding term pairs from the translated texts to be evaluated. On the one hand, term extraction methods are often tuned towards specific genres or domains (e.g. automobile, agricultural), and on the other hand they often focus on specific types of terms (e.g. MWT or NPs). We aim to evaluate how well terms are translated in students' translations on different topics from various domains. Therefore, a method of automatically identifying terms from both STs and TTs and linking them is needed.
For this purpose, in line with the prediscussed methods, we come up with a solution that uses language-independent features to train a classifier to classify ngrams into terms and non-terms in both STs and TTs, and we present a terminology-focused translation quality evaluation pipeline (See Figure 1). Our approach differs from other Machine Learning (ML) approaches based on linguistic features and context information (Li et al., 2012; Hakami and Bollegala, 2017). Instead, only minimal linguistic processing is used in our approach for data and feature set extraction, such as tokenisation and lemmatisation. The following is a brief description of the features we use to train the term classifiers.

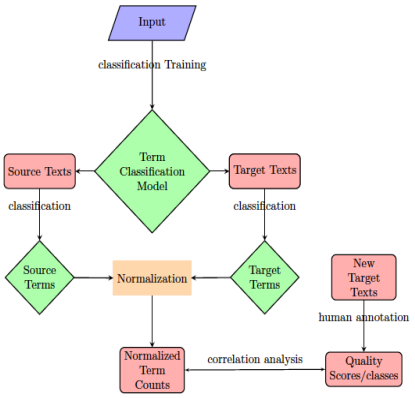

Figure 1: Terminology-focused Translation Quality Evaluation Pipeline

3. Quality Oriented Cross-lingual Term Extraction

To address the issue of cross-lingual term extraction from translational data, we present a supervised learning approach for monolingual term extraction. First, a range of representative and language-independent algorithms are exploited to compute term representations to train different classifiers. Then, monolingual terms identified by the selected, optimal classification model will be used for the normalization process, which normalizes the term counts (i.e. the number of terms 'identified by the classifier') in TTs to be the relative term frequencies in association with the number of 'terms' (as identified by the classifier as well) in STs and the length (i.e. number of tokens) of TT. This normalized term count can serve as a quality indicator in quality estimation tasks (i.e. supervised classification or regression to predict quality scores or class labels) as illustrated in the correlation analysis afterwards.

3.1. Term Classification

N-gram technique is commonly used as a language-independent approach, particularly for under-resourced language. Therefore, the term candidate classification is framed as a N-gram classification task rather than the conventional sequence labelling methods that are commonly seen in previous work (Zhou and Su, 2004; Finkel et al., 2004).
From a pragmatic point of view, our features are computed by JATE 2.0 (Zhang et al., 2016). Most representative and language-independent statistic ATR techniques are available in the package. These features (See Table 1) .
We briefly describe the features below:
TTF, namely Term Total Frequency, is the total frequency of a candidate in the target corpus. This algorithm takes into account frequency information for retrieving words or phrases that are both highly indicative of document content and highly distinctive within a text collection.
ATTF takes the average of TTF by dividing it by the number of documents in which the candidate term occurs.

Feature	Algorithm
TTF	Total Term Freqeuncy
ATTF	Average Total Term frequency
TTF-IDF	TTF with Inverse doccument Freq.
RIDF	Residual IDF
C-Value	C-Value
RAKE	Rapid Keyword Extraction
χ^2	Chi-square
Weirdness	Weirdness
GlossEx	Glossary Extraction
TermEx	Term Extraction

Table 1: Features Used for Term Extraction

TTF-IDF is adapted from the classical Term Frequency - Inverse Document Frequency (TF-IDF), which replaces the local distribution measure with global distribution across whole the corpus. It assigns higher value to words that appear more frequently in fewer number of documents across the whole corpus.

RIDF, known as residual IDF, captures the deviation of the actual IDF score of a candidate from its expected IDF score on a Poisson distribution, of which a real term (or keywords) is assumed to be higher than non-term (or ordinary words).

C-value considers the impact of frequency and length of a candidate term and thus is capable of enhancing the conventional statistics of frequency and becoming sensitive to nested terms, such as the candidate term 'T cell' nested in longer terms 'peripheral blood T cell', 'naive T cell' and 'T cell activation'.

Despite that C-value is initially proposed to extract multi-word terms (MWTs), it demonstrates flexibility to handle shorter and even single-word terms (SWTs).

RAKE, short for Rapid Automatic Keyword Extraction, can evaluate the exclusivity, essentiality and generality of extracted candidates. The measurement is based on three metrics, including word frequency, degree of word (the occurrence of a word in longer candidate MWTs and ratio of degree to frequency.

$\boldsymbol{\chi^2}$ measure is commonly used for testing whether bigram tokens co-occur by chance. JATE 2.0 adapted the measure to work with both SWTs and MWTs. If a term has no co-occurrence information, a zero score is assigned.

Weirdness, or **specificity**, is a type of contrastive ranking technique, which is particularly interesting with regard to identifying low frequent terms.

GlossEx is another hybrid approach which measures the goodness of a term by combining term specificity (i.e. termhood) and term association (i.e. unithood). The former quantifies how much an item is related to a specific domain and the latter describes the degree of association of words the term contains.

TermEx is very similar to GlossEx with extra extension of entropy-related Domain Consensus (DC) metric. DC gives more weights to a term that has even probability distribution across the documents of the domain corpus. Another two components are the Domain Pertinence (DP) and Lexical Cohension (LC), which are essentially the same as Weirdness and TC in GlossEx respectively. The final algorithm is a linear combination of the three metrics with adjustable weights (default to be 1/3 in JATE 2.0).

As mentioned earlier, all these 10 algorithms have been implemented in JATE 2.0, we just need to adapt them for working on Chinese texts.

3.2. Term Count Normalization

The normalization process aims to relate the term counts in the TTs to the terms in the STs so that the consistency of term alignments TTs can be measured and compared across different translations. Our assumption is that a higher relative number of terms counts indicates a more successful translation in terms of term adequacy which in turn contributes to the overall translation quality text-wide. The purpose of this normalization process thus is to obtain a form of term count that is comparable within translations of different lengths from STs containing different number of source terms. In the following experiment, we compute the normalized term count for each translation at the document level. Here is how the normalized term count in TTs is calculated:

$$\boldsymbol{T}_{norm} = \frac{Count_{trg} * Len_{trg}}{C_{[100]*Count_{src}}} \quad (1)$$

where T_{norm} is the normalized term count in proportion to the length of target text (Len_{trg}) in terms of the number of tokens and the number of terms in source text ($Count_{src}$), and $Count_{trg}$ is the count of terms identified in the target text, with $C_{[100]}$ a constant number 100 serving as the text length normalization base, $Count_{src}$ the number of terms in the source text.

4. Experiment

As previously stated, our experiment consists of three parts: training a monolingual term classifier, computing normalized term counts in TTs and applying the normalized term counts to quality estimation. For the last step, we do not report the results of a full quality estimation task but instead analyse the correlation of the normalized term occurrences in translations with their quality scores.

4.1. Training Monolingual Term Classifiers

4.1.1. Corpora

Five corpora, covering 3 different domains and 2 different languages (of varying sizes), are selected in the experiment to train and test our term classifiers. GENIA corpus (Kim et al., 2003) is a collection of biomedical documents and it is the most popular dataset used in ATR.

TTC, short for Terminology Extraction, Translation Tools and Comparable Corpora, a recent European project covering 8 languages, aims on the contribution of various linguistic resource for bilingual term acquisition and translation (Blancafort et al., 2010). Two English-Chinese comparable corpora (i.e. totalling 4 datasets) for two specialized domains in Wind Energy (TTC-W) and Mobile technology (TTC-M) are used in our experiment as test sets. Detailed information of all 6 corpora we used is presented in Table 2.

Corpus	# of documents	Size(tokens)	Reference Term List
GENIA	1,999	420,000	35,800
TTC-W (EN)	172	750,855	188
TTC-M (EN)	37	308,263	143
TTC-W (ZH)	178	4,263,336	204
TTC-M (ZH)	92	2,435,232	150

Table 2: Corpora Used for Training Term Classifiers

N-gram Datasets	# of terms	# of non-terms	# recall
GENIA	4,240	45,350	41%
TTC-W (EN)	120	30,925	76.5%
TTC-M (EN)	149	20,505	98%
TTC-W (ZH)	125	132,407	41.8%
TTC-M (ZH)	168	105,599	57.1%

Table 3: Terms and Non-terms in N-gram Datasets

4.1.2. Dataset Pre-processing

Firstly, all training and testing corpora are tokenised and we restrict our attention to the N-gram candidate terms with a maximum allowable length of 5 ($1 <= n <= 5$) in our current experiment. Next, stop words are removed from the list of n-gram candidates.

In the final step, training datasets and testing datasets are processed by N-gram string matching with ten features output separately by the ten algorithms. The N-gram datasets are further matched with specific Reference Term List (RTL) from each dataset. Any matched N-gram will be labelled as true positive and those having no matches will be viewed as non-terms. By this way, we eventually have 4,240 true terms from GENIA. See Table 3 for the details of our N-gram training and testing sets generated in our experiment.

4.1.3. Term Classification Models

We eventually trained 6 different models on GENIA traing corpus and then have them tested on the four TTC comparable corpora data. For both Chinese and English, we highlighted 3 optimal models each in the coloured, bold font. The classifers with best F1 score are considered as best models in our experiment. As shown in Table 4, on the Chinese test data, the optimal model achieved a precision up to 64% (true term as positive), and on the English test data, we obtained a precision up to 75%. These trained classifiers generally perform better than the Top N precisions of statistic based models (Yuan et al., 2017). Details of performances of all classifiers during the training are provided in Table 4.

4.2. Correlation with Translation Quality Scores

4.2.1. Translation Data

In the following we describe our data. The first dataset consists of 50 trainee translators' translation to a short passage about xenotransplantation (280 words). The second dataset is a course summative work from Shanghai University of International Business and Economics (SUIBE). There are 42 translations for a rotatory closure design patent in the dataset. We choose these two datasets because they are all trainee translations and they contain very domain specific words that are potentially terminology and challenging for trainees. Hereinafter, we refer to them as the XENO data and SUIBE data. The basis statistics of both datasets are shown in Table 5.

As the XENO dataset is part of our quality estimation dataset and it has been annotated by two individual annotators according to the scheme of ATA Certification Programme Rubric for Grading (Version 2011)[4].The performance of a translator is measured against four dimensions ranging from *Usefulness/Transfer* (content transfer), *Terminology/Style* (terminology and lexical equivalence), *Idiomatic Writing* (idiomaticness) to *Target Mechanics* (target language conventions), using a predefined range finder. Four subscores then make up the final score on a percentile scale.

Classifier	Testing Dataset	Precision	Recall	F1
Random Forest	GENIA(held-out)	0.80	0.84	0.8
	TTC-W(EN)	**0.79**	**0.71**	**0.75**
	TTC-M(EN)	**0.77**	**0.74**	**0.75**
	TTC-W(ZH)	**0.58**	**0.69**	**0.63**
	TTC-M(ZH)	0.57	0.60	0.58
LinearSVC	GENIA(held-out)	0.70	0.69	0.70
	TTC-W(EN)	0.66	0.79	0.72
	TTC-M(EN)	0.67	0.76	0.71
	TTC-W(ZH)	0.56	0.51	0.53
	TTC-M(ZH)	0.54	0.56	0.55
SVC RBF	GENIA(held-out)	0.73	0.73	0.73
	TTC-W(EN)	0.69	0.82	0.75
	TTC-M(EN)	**0.70**	**0.82**	**0.75**
	TTC-W(ZH)	0.51	0.53	0.52
	TTC-M(ZH)	**0.59**	**0.65**	**0.62**
MultinomialNB	GENIA(held-out)	0.64	0.59	0.61
	TTC-W(EN)	0.51	0.89	0.65
	TTC-M(EN)	0.53	0.97	0.69
	TTC-W(ZH)	0.74	0.49	0.59
	TTC-M(ZH)	**0.66**	**0.62**	**0.64**
SGD	GENIA(held-out)	0.70	0.69	0.7
	TTC-W(EN)	0.69	0.79	0.74
	TTC-M(EN)	0.67	0.82	0.73
	TTC-W(ZH)	0.60	0.49	0.54
	TTC-M(ZH)	0.58	0.59	0.58
SLR	GENIA(held-out)	0.70	0.70	0.70
	TTC-W(EN)	0.68	0.81	0.74
	TTC-M(EN)	0.70	0.81	0.75
	TTC-W(ZH)	0.58	0.51	0.54
	TTC-M(ZH)	0.59	0.59	0.59

Table 4: Model Performance on Development and Testing Datasets

Dataset	Domain	Passages	# of sentence	Length
XENO	Xenotransplantation	50	14	234 ~ 473
SUIBE	Patent	42	11	297 ~ 376

Table 5: Basic Statistics for Two Trainee Translation Datasets

4.2.2. Query Terms in Translations

As the list of terms is first generated in the term classification process at the corpus level, we need to query the identified terms in translation case by case. Meanwhile, in order to mitigate the influence of their negative effects, we adopted a tri-gram (letter for English and character for Chinese) similarity matching policy on any candidate term pairs. If any ST or TT ngram has a similarity larger than 0.7 with the candidate terms identified by the classifier from the whole lot of ST or TTs, we deem them a success term

[4] http://www.atanet.org/certification/aboutexams_rubic.pdf

		ST			TT		
		precision	recall	F1	precision	recall	F1
XENO	SLR	0.19	0.5	0.28	0.01	1.00	0.01
	MNB	0.02	0.88	0.04	0.00	0.88	0.01
	RF	0.04	0.5	0.07	0.01	1.00	0.01
	SGD	0.18	0.5	0.26	0.01	1.00	0.01
	SVCRBF	0.14	0.38	0.20	0.01	1.00	0.02
	LinearSVC	0.19	0.5	0.28	0.02	0.75	0.04
SUIBE	SLR	0.47	0.47	0.47	0.00	0.94	0.01
	MNB	0.26	0.41	0.32	0.01	0.88	0.01
	RF	0.18	0.18	0.18	0.00	0.94	0.01
	SGD	0.47	0.47	0.47	0.00	0.94	0.01
	SVCRBF	0.41	0.41	0.41	0.02	0.94	0.03
	LinearSVC	0.47	0.47	0.47	0.01	0.88	0.02

Table 6: Monolingual Terminology Identification on Two Datasets

translation. Therefore, terms, such as 'slightly conical pipe segments (' and 'conical pipe segment' and 锥形 管 (6 and 圆锥形 管 段 are likely to be matched when we are going to find out how many terms are correctly translated.

4.2.3. Evaluation

To investigate whether the automatically identified terminology are related to the quality of the trainees' translations, we compute the Pearson correlation coefficient (r), Spearman rank correlation coefficient (R_s) and Kendall's Rank Correlation Coefficient (τ) (Bolboaca and Jäntschi, 2006). The correlation coefficients are calculated as:

$$r = \frac{n\sum x_i y_i - (\sum x_i \sum y_i)}{\sqrt{n\sum {x_i}^2 - (\sum x_i)^2}\sqrt{n\sum {y_i}^2 - (\sum y_i)^2}},$$

where n is the number samples and x_i,y_i are the paired instances of the observed and estimated variables,

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)},$$

where n is the number of smaples and d is the pairwise distances of the ranks of the variables x_i and y_i, and

$$\tau = \frac{S}{\sqrt{n(n-1)/2 - T}\sqrt{n(n-1)/2 - U}}$$

$$T = \sum_t t(t-1)/2$$

$$U = \sum_u u(u-1)/2,$$

where S is the difference between the number of concordant pairs[5] and the number of discordant pairs[6], t is the number of observation of variable x that are tied[7] and u is the number of observation of variable y that are tied[8].

4.3. Results and Findings

We report the confusion matrix of terms identified monolingually by the six classifiers trained above from the English STs and their Chinese TTs in Table 6.

On the XENO data, as is shown in Table 6, six classifiers except for the Multinomial Bayes (MNB) perform rather consistently on the English source text, but display rather apparent variation on the Chinese translations. We manually analysed their predictions on the termhood of ngrams. Almost all classifiers have successfully identified 'xenotransplant', 'xenotransplantation','transplant surgeon' as terms, but they failed to identify two terms 'recipient' (受体) and the institute 'America's Food and Drug Administration'. A possible explanation is that they are ignored because they are singletons which pose difficulty for our statistics-based features to capture the subtlety. For the Chinese translations, all the representative terms, such as '器官 移植 医生' (transplant surgeon), '异种 器官 移植 手术' or '异种 器官 移植' (xenotransplantation), have been successfully identified.

Meanwhile, on the SUIBE patent data, classifiers manifested a significant deterioration of performance, with many terms mistagged. The majority of terms in the ST, such as 'rotary closure', 'neck inner wall', 'radial rib', 'conical pipe segment' and 'pivoting range', were misclassified. On Average, only less than one-third of the true terms (22) from the ST could be recognized by our classifiers.This is in contrast to the good recall of the classifiers on the translations, as is shown in the confusion matrix in Table 6. Term equivalents in the translations for those ST terms that were misclassified are able to be identified by our classifiers (together with a large proportion of false positives). This huge drop of performance on SUIBE ST may be due to the cross-domain effect. As we directly apply the classification model trained from the biomedical data (GENIA 1.0), though our features are supposed to be domain independent, we suspect the domain-shift issue may still impact the classification model. Note that in the table low precision of term classification may be due to the ngram generation process that produces a large amount of sequences of words that are classified as terms, recall is more important in our study though.

As for the alignment process, we report the types, e.g. one-to-one[9], one-to-many[10], many-to-one [11], many-to-many[12] and the length of alignment information for the ST and TTs. Judging from the list of the aligned pairs, the process of alignment has introduced some noises. Either ST terms or equivalent term translations often contain extra words, punctuations. For instance, The pair 'xenotransplantation.' and (异种 器官 移植) has an extra full stop mark. Other types of errors that could be problematic for the later term query include one-to-many and many-to-one alignment, which will cause confusion to the query for matching terms in both ST and TTs. In order to mitigate the influence of their negative effects, we adopted a tri-gram (letter for En-

Data	Length		Type	
XENO	1-word	143	one-to-one	42
	2-word	282	one-to-many	101
	3-word	341	many-to-one	678
	4-word	621	many-to-many	4833
PATENT	1-word	71	one-to-one	57
	2-word	127	one-to-many	14
	3-word	113	many-to-one	50
	4-word	80	many-to-many	334

Table 7: Alignment Types and Distribution

[5]For any pair of observations (x_i, y_i) and (x_j,y_j), where $i \neq j$,they are concordant if the ranks for both elements agree.

[6]if $x_i > x_j$ and $y_i < y_j$, or $x_i < x_j$ and $y_i > y_j$.

[7]if $x_i = x_j$

[8]if $y_i = y_j$

[9]one ST word is aligned to one TT word.

[10]one ST word is aligned to more than one TT word.

[11]More than one ST word is aligned to one TT word.

[12]More than one ST word is aligned to more than one TT word.

Dataset	Terms	Human Annotation	Correlation		
			Pearson	Spearman	Kendall Tau
XENO	$\mu = 3.68, SD = 3.45$	Usefulness/Transfer ($\mu = 24.31, SD = 4.73$)	$r = 0.43, p < 0.01$	$\rho = 0.48, p < 0.0001$	$\tau = 0.37, p < 0.0001$
		Terminology/Style ($\mu = 16.67, SD = 3.06$)	$r = 0.46, p < 0.01$	$\rho = 0.52, p < 0.0001$	$\tau = 0.39, p < 0.0001$
		Idiomatic Writing ($\mu = 17.12, SD = 2.63$)	$r = 0.32, p = 0.02$	$\rho = 0.35, p = 0.01$	$\tau = 0.26, p = 0.01$
		Target Mechanics ($\mu = 9.79, SD = 1.35$)	$r = 0.36, p = 0.01$	$\rho = 0.39, p < 0.01$	$\tau = 0.31, p < 0.01$
		Final Score ($\mu = 71.57, SD = 12.41$)	$r = 0.66, p < 0.0001$	$\rho = 0.72, p < 0.0001$	$\tau = 0.55, p < 0.0001$
SUIBE	$\mu = 10.52, SD = 10.19$	Final Score ($\mu = 87.07, SD = 5.86$)	$r = 0.53, p < 0.001$	$\rho = 0.60, p < 0.001$	$\tau = 0.44, p < 0.0001$

Table 8: Correlation between Term occurrences and Translation Quality

glish and character for Chinese) similarity matching policy on any candidate term pairs. If any ngram has from a ST and a TT both have a similarity larger than 0.7 with the candidate term pairs in the aligned list, we deem them a success term translation. Therefore, terms, such as 'slightly conical pipe segments (' and 'conical pipe segment' and 锥形 管 (6 and 圆锥形 管 段 are likely to be matched when we are going to find out how many terms are correctly translated.

According to the values of three correlation metrics in Table 8, for the XENO dataset, the number of terms identified in both datasets show a positive linear relationship with the four subscores (See Table 8) inbetween weak and moderate ($p < 0.01$). In contrast, the occurrence of terms with the final score (weighted summation of all subscores) goes up beyond moderate ($p < 0.0001$). For the PATENT data, as it has only one final score for all translations, we could also find a moderate linear relationship between the rightly translated terms in the translations ($p < 0.001$). Despite two datasets are evaluated by different annotators under various criteria, correlation scores, either Pearson r, Spearman ρ or Kendall's τ all suggest that the number of correctly translated terms does contribute to translation quality on the whole.

However, it is surprising that there exists only a weak correlation between the second subscore (Terminology/Style) and the term occurrence in the translations. We checked those translations with zero hit of terms but over strong quality scores. We found translation of terminology, semantic adequacy and language fluency are present in the translation indeed, see Table 9. Typical terms in the specific domain, such as '异种 器官 移植 '(xenotransplantation), '器官 移植 外科 医生'(transplant surgeons),'美国 食物 药物 管理局'(America 's food and drug administration) are adequately translated. One thing in common with these translations is that through the translation terms are rendered with slight variation. For example, in one sample, both '器官 移植 外科 医生'(transplant surgeons) and '器官 移植 手术 师'(transplant surgery technician) are used for the same source term 'transplant surgeon'. Both translations are acceptable expressions in Chinese in terms of adequacy and fluency. This term inconsistency or variation may have to do with why such translations are evaluated reasonably high even with few or no term counts by our trained term classifiers. It implies that our approach of term classification may have fallacy in handling term variation.

#	ST	TT
1	Transplant surgeons work miracles.	器官移植外科医生带来了奇迹。
	They take organs from one body and integrate them into another,	他们将器官从一个身体中取出并将它们植入他者体内,
	granting the lucky recipient a longer, better life.	让那些幸幸得到 它们的人活得更长，更好。
2	America's food and drug administration has already published draft guidelines for xenotransplantation.	美国的食品药物管理局已经出版了异种器官移植 草案 细则。
	The ethics of xenotransplantation are relatively unworrying.	这种手术在伦理道德 领域 相对而言，不那么令人担忧了。
3	So far attempts to make artificial organs have been disappointing:	到目前为止，试图人工制造器官的可能性已经被否定了，
	Nature is hard to mimic. hence the renewed interest in trying to use organs from animals.	毕竟自然是很难去模仿的，因此，人们正将更多的目光集中在动物器官上。

Table 9: Adequately Translated Terms:Term Variation

5. Conclusion

In this study, we explored ways of identifying terms from monolingual texts and integrate them into investigating the contribution of terminology to translation quality. It is found that the number of term frequencies identified automatically has weak linear correlation with the four subscores for the xenotransplantation data. When it comes to the overall final score for both datasets, such correlations increase to be above moderate and strong. This study indicates that the term occurrence in translation could be an valuable quality indicator for estimating translation quality. In the future, we deem it necessary to try other weakly supervised method to improve term identification accuracy, particularly those low frequency terms and their variations. Ultimately, term occurrence will be incorporated into the existing feature set (Yuan et al., 2016) in quality estimation tasks.

6. Acknowledgements

We are thankful to the generosity of Dr. Xiaoying Hu's sharing the SUIBE data and Jie Gao's kind assistance in preparing the monolingual training and testing data. The first author is supported by the China Scholarship Council-University of Leeds Scholarship (Support No. [2013]3009).

7. References

Blancafort, H., Daille, B., Gornostay, T., Heid, U., Méchoulam, C., and Sharoff, S. (2010). TTC: Terminology Extraction, Translation Tools and Comparable Corpora. In European Association for Lexicography, editor, *Proceedings of the XIV Euralex International Congress*, pages 263–268, Leeuwarden/Ljouwert, Netherlands, July.

Bolboaca, S.-D. and Jäntschi, L. (2006). Pearson versus spearman, kendall's tau correlation analysis on structure-activity relationships of biologic active compounds. *Leonardo Journal of Sciences*, 5(9):179–200.

Brunette, L. (2000). Towards a terminology for translation quality assessment: A comparison of tqa practices. *Translator: Studies in Intercultural Communication*, 6(2):169–182.

Erdmann, M., Nakayama, K., Hara, T., and Nishio, S. (2009). Improving the extraction of bilingual terminology from wikipedia. *ACM Trans. Multimedia Comput. Commun. Appl.*, 5(4):31:1–31:17, November.

Finkel, J., Dingare, S., Nguyen, H., Nissim, M., Manning, C., and Sinclair, G. (2004). Exploiting context for biomedical entity recognition: from syntax to the web. In Nigel Collier, et al., editors, *COLING 2004 International Joint Workshop on Natural Language Processing*

in Biomedicine and Its Applications (NLPBA/BioNLP) 2004, pages 91–94. COLING, August.

Gaizauskas, R., Paramita, M. L., Barker, E., Pinnis, M., Aker, A., and Solé, M. P. (2015). Extracting bilingual terms from the web. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 21(2):205–236.

Gómez Guinovart, X. and Simoes, A. (2009). Parallel corpus-based bilingual terminology extraction. In Anne Condamines et al., editors, *8th International Conference on Terminology and Artificial Intelligence*, Toulouse,France, November.

Hakami, H. and Bollegala, D. (2017). A classification approach for detecting cross-lingual biomedical term translations. *Natural Language Engineering*, 23(1):31–51.

Hartley, A., Mason, I., Peng, G., and Perez, I. (2004). Peer and self-assessment in conference interpreter training. research project. pedagogical research fund in languages, linguistics and area studies. *The Higher Education Academy, Heslington, United Kingdom. Retrieved May*, 7:2005.

Karoubi, B. (2016). Translation quality assessment demystified. *Babel*, 62(2):253–277.

Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). Genia corpus–a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl 1):i180–i182.

Kim, T., Hwang, M., Hwang, M., Song, S., Jeong, D., and Jung, H. (2015). Translation of technical terminologies between english and korean based on textual big data. *Software: Practice and Experience*, 45(8):1115–1126.

Li, L., Dang, Y., Zhang, J., and Li, D. (2012). Domain term extraction based on conditional random fields combined with active learning strategy. *Journal of Information &Computational Science*, 9(7):1931–1940.

Lommel, A., Uszkoreit, H., and Burchardt, A. (2014). Multidimensional quality metrics (mqm). *Tradumàtica*, No.(12):0455–463.

Macken, L., Lefever, E., and Hoste, V. (2013). Texsis: Bilingual terminology extraction from parallel corpora using chunk-based alignment. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 19(1):1–30.

Rocheteau, J. and Daille, B. (2011). TTC TermSuite: A UIMA Application for Multilingual Terminology Extraction from Comparable Corpora. In *5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 9–12, Chiang Mai, Thailand, November. System Demonstrations.

Secară, A. (2005). Translation evaluation–a state of the art survey. In *Proceedings of the eCoLoRe/MeLLANGE Workshop*, pages 39–44, Leeds, UK.

Warburton, K. (2013). Processing terminology for the translation pipeline. *Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication*, 19(1):93.

Xu, R. and Sharoff, S. (2014). Evaluating term extraction methods for interpreters. In *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)*, pages 86–93, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Xu, Y., Chen, L., Wei, J., Ananiadou, S., Fan, Y., Qian, Y., Eric, I., Chang, C., and Tsujii, J. (2015). Bilingual term alignment from comparable corpora in english discharge summary and chinese discharge summary. *BMC bioinformatics*, 16(1):149.

Yuan, Y., Sharoff, S., and Babych, B. (2016). Mobil: A hybrid feature set for automatic human translation quality assessment. In *Proc Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia, May.

Yuan, Y., Gao, J., and Zhang, Y. (2017). Supervised learning for robust term extraction. In *Proceedings of 2017 International Conference on Asian Language Processing (IALP)*, Singapore, December. IEEE.

Zhang, Z., Gao, J., and Ciravegna, F. (2016). Jate 2.0: Java automatic term extraction with apache solr. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Zhou, G. and Su, J. (2004). Exploring deep knowledge resources in biomedical name recognition. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications*, JNLPBA '04, pages 96–99, Geneva, Switzerland. Association for Computational Linguistics.