Investigating the Influence of Bilingual Multiword Units on Trainee Translation Quality

Yu Yuan[†] [‡], **Serge Sharoff**[†]
[‡] School of Languages & Cultures, Nanjing University of Information Science & Technology, 210044, China
[†] Centre for Translation Studies, University of Leeds, LS2 9JT, United Kingdom
hittle.yuan@gmail.com, s.sharoff@leeds.ac.uk

Abstract

We applied a method for automatic extraction of bilingual multiword units (BMWUs) from a parallel corpus in order to investigate their contribution to translation quality in terms of adequacy and fluency. Our statistical analysis is based on generalized additive modelling. It has been shown that normalised BMWU ratios can be useful for estimating human translation quality. The normalized alignment ratios for BMWUs longer than two words have the greatest impact on measuring translation quality. It is also found that the alignment ratio for longer BMWUs is statistically closer to adequacy than to fluency.

Keywords: Bilingual multiword unit, normalised alignment ratio, translation quality estimation, trainee translators

1. Introduction

The overall goal of this study is to investigate automatic quality assessment for translators, especially for the trainees. Many problems related to quality stem from inappropriate translation of MWUs, such as idioms (e.g. *barking at the wrong tree*, *add insult to injury*, etc.), terms (e.g. *recursive function*, *closed captioning*, etc.), phrasal verbs (e.g. *fall out*, *give up*, etc.), or named entities (e.g. *Vice-Chancellor Sir Timothy Michael Martin O'Shea*, *New York City*, etc.). Baker (2011, pp:72-73) identified four types of errors when translating idiomatic expressions: 'no equivalent', 'similar counterpart but different context', 'idiom play', 'different conventions of using idioms'. A trainee translator in our dataset produced an unacceptable calque of *as brave as a lion* into Chinese 勇猛如狮, while the correct idiom should have been 勇猛如虎 'as brave as a tiger'. This erroneous treatment is what Baker terms 'different conventions of using idioms'.

Based on observations of source language (SL) idioms and their Persian translation, researhers have found that explicit loss,implicit loss, modified loss and complete loss are common resultant categories of cultural losses (Zebardast and AbuSaeedi, 2015; Zebardast, 2015). Translation of MWUs is problematic because it is associated with cultural issues (Min, 2007), systematic language variation (Wang and Nian, 2004) and the translator's failure to decipher the meanings of MWUs in question (Abu-Ssaydeh, 2004). Therefore, this paper assesses the contribution of translating MWUs to the overall translation quality, particularly for trainee translations.

In this paper we investigate:

1. how can we extract BMWUs for trainee translation quality evaluation?
2. how is the use of BMWUs related to the quality of trainee translations?
3. How do BMWUs of different lengths contribute to the trainee translation quality?

The term BMWUs refers to recurrent sequences of words that are translations to each other in aligned bilingual texts. In other studies they are also known as (bilingual) phrase alignments, aligned phrases units (sequences), bilingual alignments, bilingual phraseological units, bilingual phraseology, etc. In this paper, we use the term bilingual multiword units (BMWUs) hereinafter.

We start with the automatic acquisition of BMWUs from parallel corpora and use them as base against which to query and compare trainee translations. Our main contribution is designing and using BMWU-related features as quality indicators for human translation quality estimation.

2. Methodology

In this section, we will describe how BMWUs are used in this study, the method we use to extract BMWUs and our exploration of the relationship between BMWUs of different lengths and translation quality in terms of fluency and adequacy using the mixed-effect modelling.

2.1. Extraction of BMWUs

As we work with the parallel corpora we can combine the task of identification of monolingual MWUs with the process of bilingual alignment, saving us from the trouble of a difficult task of monolingual MWU identification (Sag et al., 2002). The alignment process which is aimed at producing phrase tables for statistical machine translation (SMT) (Koehn et al., 2003) can be based on flat models or on hierarchical models. In traditionally used flat IBM family models, the phrase tables are generated in two steps, first generating word alignments and then extracting a scored table of phrase pairs. However, this often yields a large proportion of unwanted word alignments, as there are only minimal phrases memorized by the model(DeNero and Klein, 2008), so it has to be combined with heuristic phrase extraction to exhaustively combine adjacent phrases permitted by the word alignment (Och et al., 1999).

In contrast, Bayesian-based phrase alignment as proposed in (Neubig et al., 2011) is a model for joint phrase alignment and extraction using non-parametric Bayesian methods and inversion transduction grammars (ITGs). A hierarchical ITG model relies on the Pitman-Yor process (Pitman and Yor, 1997) to directly use probabilities of the model as

Table 1: Professional Translations for the Same Source Text Phrase

Source	Translation	Frequency
总而言之	in sum	4
	in summary	4
	all in all	3
	in short	5
	in conclusion	2
	in general	4
	overall	4
	$--$ (omitted)	2

a replacement for the phrase table generated by heuristic techniques, e.g. intersection, grow-diag in Giza $++$ (Och and Ney, 2003). Because of its compactness and competitive accuracy, we choose this method over other standard heuristic alignment tools (e.g. Giza $++$, fast_align (Dyer et al., 2013), etc.) to obtain an aligned list of MWUs.

Our study is mainly focused on investigating the contribution of BMWUs of varying lengths to the translation quality of trainee translations. For this task, we need an authentic database of BMWUs from a sizeable bilingual corpus of professional translations to have enough statistics for MWU identification and pruning. The extracted BMWUs will be used to measure the degree of adequacy and fluency of trainee translations. If professional translators view language expressions that can transfer meaning unambiguously as the basic translation units (Baobao et al., 2002), we believe, such a database of BMWUs can be a useful resource for human translation quality estimation. Investigation of BMWUs of different lengths can be viewed as part of the feature engineering for human quality estimation task.

It is often observed that human translators translate group of words as a whole and words are rarely treated as the working translation units individually. Variation in translations from a large corpus and distribution of frequencies, as illustrated in Table 1[1] will lead to varying probabilities of the aligned BMWUs. Though the list of candidate translations is not exhaustive, if translations by students are not in this list, they are more likely to be inappropriate translations.

Therefore, our working hypothesis is that the consistency of BMWUs in trainee translations in relation to professional translations can be interpreted as a higher degree of semantic adequacy and stylistic fluency. In the following experiment, we will use the Bayesian-based ITG method to automatically extract phrasal alignments of different lengths (1-4 words) from a large parallel corpus, and then compute the normalized ratios of these aligned phrasal sequences for each trainee translation at the document level. Here is how the normalized ratio of BMU is calculated:

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

where R_{norm} is the normalized ratio of BMWUs in proportion to the length of source text (Len_{srt}) in terms of the number of tokens, and $Count_{trg}$ is the count of BMWUs in the target text, with $C_{[100]}$ a constant number serving as the normalization base [2]. Note that in our calculation the ratio is computed in relation to the source text length. Therefore, the returned ratio is equivalent to the recall of BMWUs. This implies we can also compute the precision of BMWUs in relation to the target text length.

2.2. Multilevel Mixed Effects Modelling

Multilevel mixed effects modelling is a way of understanding the types of relationships you can examine or consider of your data. This technique of data analysis is especially useful when observations at one level of analysis are nested within observations at another, and when other categorical or hierarchical data types are involved. In our case, we are investigating how normalized ratios of BMWUs of varying lengths contribute to translation adequacy and fluency. Thus, our data (normalized ratios of BMWUs) are nested in a different category, e.g., BMWU length.

It has advantages of analysing phenomena at multiple levels simultaneously and identifying important relationships across different levels of analysis. The recent development of statistical modelling has made it feasible to analyse group and individual variance simultaneously. This modelling is based on a generalized linear model with a linear predictor involving a sum of smooth functions of covariates (Wood, 2017, pp:161). In general, the model has the form :

$$g(\mu_i) = \mathbf{A}_i\gamma + \sum\nolimits_i f_j(x_{ji}), y_i \sim \mathbf{EF}_{(\mu_i,\phi)} \quad (2)$$

where $\mathbf{A}_i$ is the i^{th} row of a parametric model matrix, with corresponding parameters γ, f_j is a smooth function of covariate x_j, and $\mathbf{EF}(\mu_i, \phi)$ denotes an exponential family distribution with mean μ_i and scale parameter ϕ (Wood, 2017, pp:249). This modelling technique allows for the flexibility and convenience of specifying the model in terms of smooth functions, which can then be estimated form data using cross validation or marginal likelihood maximization.

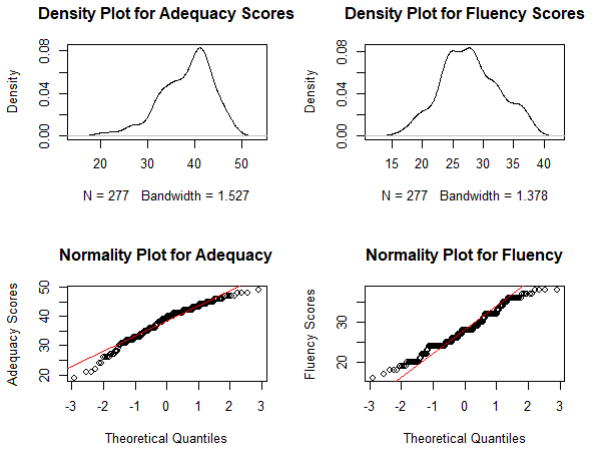
In this study, a mixed-effect modelling is carried out, aiming to explore the significance of BMWUs accounting for students' translation scores in adequacy and fluency and how BMWU alignments of different lengths impact the quality scores, when there are multiple levels of translation quality and alignment lengths are involved.

The response variables in our study are the adequacy and fluency scores, and the explanatory variables in our experiment are the normalized alignment ratios (see Equation (1)), coded as NrmALR, and MWU length information, coded as AL1, AL2, AL3, in the translations. Meanwhile, in order to answer the questions listed above, we are also interested in knowing the interaction between the BMWU length and their normalized ratios. Thus, in our design, we keep these two as the fixed effects and the training corpus size (two sizes) (coded as TrCorpS1, TrCorpS2) and each individual translation as the random effects.

[1] Generated from http://www.linguee.com

[2] The setting of the normalization base depends on the average length of the trainee translations, and in our case they are around $2 - 400$ words. Hence, we set 100 as the base for normalization.

Figure 1: Normality Test for Adequacy and Fluency Scores



Both adequacy and fluency scores are not normally distributed (Shapiro-Wilk normality tests $p < .05$), as shown in Figure 1, Adequacy scores are left-skewed due to central tendency towards upper bound (more scores above the third quartile), and Fluency scores are short tailed, indicating very small variances between them.

To proceed, we converted the adequacy scores and fluency scores into ordered categorical data (Liu and Agresti, 2005) so that it fits the condition that the expected value of the response variable (e.g. adequacy, fluency) follows a logistic distribution. Analysis was performed in the R package mgcv[3] for generalized additive mixed modelling.

We treat the normalized alignment ratio and phrase alignment length as main effects and several combinations of random effects of training corpora and sample IDs, using the ordered categorical data distribution family. This family of method is for use with generalized additive model, implementing regression for data following a logistic distribution. The observed categories are coded $1, 2, 3 \ldots$, up to the number of categories (Wood et al., 2016).

3. Parallel Corpus and the Trainee Data

For this study we use the English Chinese parallel UM corpus of mixed domains (Tian et al., 2014). It is a multi-domain and balanced parallel corpus covering several topics and text genres, including education, law, microblogs, news, science, spoken, subtitles and theses. The English part is tokenised with the scripts included within the Statistical Machine Translation system moses (Koehn et al., 2007). The Chinese part is segmented with Jieba Chinese word segmentation module.[4]

As for trainee translations, we have 277 student translations in six different domains scored by two raters 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, so that the total score of a student translation can be in the [0-100] range (Table 3).

Table 2: Statistics of UM Parallel Corpora

Domains	Languages	Tokens	Average Length	Vocabularies	Sentences
News	English	8,646,174	19.21	274,546	45,000
	Chinese	15,277,414	33.95	47,902	
Spoken	English	1,836,670	8.35	107,923	220,000
	Chinese	3,033,052	13.79	9,011	
Laws	English	5,926,316	26.94	66,330	220,000
	Chinese	8,783,941	39.93	14,723	
Thesis	English	5,962,590	19.88	378,679	300,000
	Chinese	10,514,430	35.05	149,110	
Education	English	8,401,095	18.67	293,595	450,000
	Chinese	13,749,570	30.56	38,663	
Science	English	598,050	2.22	115,968	270,000
	Chinese	1,527,849	5.66	8,927	
Subtitles	English	2,299,742	7.67	101,423	300,000
	Chinese	3,818,490	12.73	13,854	
Microblog	English	72,144	14.43	12,083	5,000
	Chinese	125,415	25.08	3,525	
Total	English	33,742,781	13.29	832,518	2,215,000
	Chinese	56,830,161	22.51	209,729	

Table 3: Basic Statistics of English-Chinese Translational Data

Source Text	Domain	Topic	Statistics			
			Source Text		Translation	
			# of Sentences	# of words	# of sentences (mean)	# of words (mean)
ST1	Science fiction	Insects	11	261	10	317
ST2	Scoial life	Marriage	15	259	14	311
ST3	Sports	Walking	13	289	12	353
ST4	Short story	Preseverance	15	313	14	410
ST5	Literature	Essayist	5	229	4	246
ST6	Science	xenotransplantation	13	266	11	372

Table 4: Range Finders for Different Grades of Translation

Grades	Usefulness/transfer	Terminology/style	Idiomatic Writing	Target Mechanics
Standard	29-35	21-25	21-25	13-15
Strong	22-28	16-20	16-20	10-12
Acceptable	15-21	11-15	11-15	7-9
Deficient	8-14	6-10	6-10	4-6
Minimal	1-7	1-5	1-5	1-3

Two Chinese native annotators, both are PhD students in Translation Studies, following the scoring scheme of ATA Certification Programme Rubric for Grading (Version 2011),[5] measure the performance of a translator against four dimensions ranging from Content Transfer (CT), terminology (T), idiomatic writing (I) and target language conventions (TC), see Table 4. They evaluated the translations based on the degree to which learner translators have transferred the meaning completely (combining CT and T into the Adequacy score) and followed the rules and conventions of the target language (I and TC combined into the Fluency score). The inter-annotator agreement is substantial (Krippendorff's $\alpha = .77$ for Adequacy and .89 for Fluency).

[3] https://cran.r-project.org/web/packages/mgcv/index.html

[4] https://github.com/fxsjy/jieba

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

Table 5: Length and Type Distribution of BMWUs

Length	Counts	Alignment Types	Counts
1-word	648,611	one-to-one	374,840
2-word	1,835,261	one-to-many	2,73,771
3-word	1,889,009	many-to-one	362,849
4-word	1,344,276	many-to-many	4,705,697
Total	5,717,157	Total	5,717,157

Table 6: BMWU Alignment Accuracy (threshold DTP >= 0.2)

Alignment	Top1000(%)	Bottom1000(%)
BMWU(Single word)	96.3	26.9
BMWU(2-more words)	98.7	97.2

Table 7: An Excerpt of Aligned BP Extracted

BP(source)	BP(target)	DTP	ITP
tax credits	税额 减免	0.37	0.25
	税款 抵减	0.52	0.34
	税收 减免	0.08	0.05
	的 税收 优惠	0.90	0.05
	税收 优惠	0.05	0.05
	税额 抵免	1.00	0.05
	税收 抵免	0.59	0.17
and tax credits	或 减免	0.50	0.50
	和 税 减免	1.00	0.50
as tax credits	诸如 税收 扣减	1.00	1.00
continuous tax credits	持续性 税收 扣除	1.00	1.00
for tax credits	享受 税额 减免	1.00	1.00
investment tax credits	投资 税额 减免	1.00	1.00
in tax credits	税收 抵免	0.02	1.00
production tax credits	生产 税额 减免	0.33	1.00

4. Findings and Discussion

4.1. Considerations of BMWU Alignment Quality

We extracted BMWUs of a maximum length of 4. The reason we did not go up to longer sequences of phrases is that longer alignments (> 4 words) are very rare in students' translations. We include one word MWUs to account for one-to-many and many-to-one alignments, i.e. some MWUs are produced as translations of a single source word, and single words are translations of source MWUs. We eventually have 9.63 million pairs of phrasal alignments without exclusion and 5.72 million pairs after filtering per direct translation probability (DTP) and inverse translation probability (ITP), the selected thresholds of DTP and ITP $>= 0.01$, See Table 5.

Longer MWU alignments from the training corpus are generally more accurate than shorter ones. Accuracy of BMWU alignments remain stable even though their direct translation probabilities decrease. In order to evaluate the validity of entries in the extract list of BMWUs, we set the DTP threshold for all alignments to be 0.01 and then sort them per their DTP values and discard all non-English-Chinese pairs, i.e., Null alignments, punctuations, symbols, strings alignments and English-English alignments. That is to say, our evaluation is based on English-Chinese alignment pairs only. We compared the top 1000 and bottom 1000 entries for single word alignments and phrasal alignments (BMWUs of 2 to more words). See Table 6 for details. We believe that the reason one-word alignments contain many false matches is because of word segmentation, e.g. 无用功 ('unproductive work') should be matched by two English words to be equivalent, and word associations in the context, e.g. 拖垮 ('drag down') should be the cause of lack of productivity and it is supposed to occur within the near context. Nevertheless, as Table 6 shows, these BMWUs, particularly those longer than one word, when selected as bilingual correspondences with reasonably good accuracy, can be readily usable.

In preparing the dictionary of bilingual phrase alignments for query their occurrences in trainees' translations, four-word alignments are eventually discarded because their extremely low frequencies in the target translations. Analysis of the phrase alignment dictionary will be presented in the final paper. Note that while DTP may be a useful signal of alignment certainty we cannot take it for granted that lower probabilities nullify the legitimate translation equivalents, as many pairs with very low probabilities are valid translations to each other, for instance 报价和目录 ('quote and directory') and *quotation and catalogues* are aligned at a probability of 0.08 but they are clearly valid alignments. This again explains why in Table 6 there is no significant difference for longer BMWUs ranked by DTPs (top and bottom). Also this shows that we need keep good coverage of BMWUs. For this study, we set the threshold of direct translation probability at 0.02. This cutting-off value eventually allows us to have 3.5 millions pairs of bilingual phraseological pairs, a much slimmer list of phrase table, as illustrated in Table 7. In the Table, the third and fourth column are direct (conditional) translation probabilities (DTP, i.e. translation probability from English to Chinese) and inverse translation probabilities (ITP, i.e. translation probability from Chinese to English).

With this acquired probabilistic dictionary of BMWUs, each 1- word, 2-word and 3-word lexical unit in the TT is then queried against each trainee translation to find the matches between the corresponding units. When calculating the translation probability of a TT, the parallel sentences of ST and TT are taken as a unit, matched against the bilingual lexicon trained from last step and the results is the logarithm of all matched translation probabilities in each translation. Normalized BMWU ratios of varying lengths are calculated as per Equation (1). The rationale for this score is these four indexes come from the conception that good translations are often close to professional or expert translations and therefore direct translation probabilities trained from large parallel corpora of professional translations can be a good criterion for the word choices by translators. True alignments of higher probabilities are more likely to be the right candidates and in this way, for a translation, Therefore, there would be variation of the normalized phrase alignment ratios across translations of different quality. Better translations would have higher normalized ratios of aligned BMWUs than inferior ones. For the later modelling, translations are divided into three levels of quality groups, i.e. poor ([39, 54.7]), average ([54.7, 70.3]) and good ([70.3, 86]) per their final scores on ade-

Table 8: Illustration of Automatically Aligned Phrases(1-3 words)

Phrase Length	English Phrases	Chinese Phrases	DTP	Gloss
1-word	offences	所犯	0.33	committed by
		犯罪行为	0.29	crimes
		而犯	0.44	commit
		系由该	1.00	This is committed by
		犯罪	0.22	commit a crime
2-word	traffic offences	违犯 交通法规	0.91	violation of traffic regulations
		交通 罪行 的	1.00	traffic offence (about)
		交通 罪行	1.00	traffic offence
3-word	international drug trafficking	国际 贩毒	0.39	international drug trafficking
		国际 毒品 贩运	0.86	international drug trafficking
		国际 药物 贩运	1.00	international medicine trafficking

quacy and fluency. The BMWU ratio for each translation is further divided into three length groups, i.e. one, two and three word alignment ratio.

4.2. Mixed Effects Modelling

A model is eventually selected per Akaike's Information Criterion (AIC=4287.18) using a restricted maximum likelihood as our final model for adequacy. Results are reported in Table 9 below. The random effect of interaction between specific sample and the normalized alignment ratio (NrmALR) is also significant (${\chi_{(246.7,276)}}^2 = 2214$, $p < .0001$). The model has revealed that longer BMWUs contribute more to the adequacy scores

Table 9: Fixed Effects of Alignment Ratio and Alignment Length on Adequacy

Response Variable	Effect	Estimate	SD	Z	$Pr. > t$ or Z	AIC
Adequacy	Intercept	-4.91	0.53	-9.14	< 0.0001	4287.18
	NrmALR	4.35	0.39	11.15	< 0.0001	
	AL2	3.19	0.69	4.64	< 0.0001	
	AL3	6.52	0.73	8.82	< 0.0001	
	TrCorpS2	-0.67	0.09	-7.70	< 0.0001	
	NrmALR:AL2	5.04	0.48	10.59	< 0.0001	
	NrmALR:AL3	11.22	0.75	14.93	< 0.0001	

The selected model tells that, ignoring other variables, the normalized BMWU alignment ratio has a significant impact on the adequacy scores (${\chi_{(2)}}^2 = 314.37$, $p < .0001$; $z = 11.15$, $p < 0.0001$), and similarly, three word alignments (AL3) and two word alignments (AL2), in contrast to one word, have a significant influence on adequacy scores. In addition, there is a significant interaction between two word alignment (AL2) and the three word alignment (AL3) and the normalized BP alignment ratio, which suggests the adequacy scores across different alignment length groups (specifically, longer than one word BMWs) are significantly different in terms of the normalized alignment ratio.

We fitted three mixed effect model with the same method using fluency scores as the response variable. Significant random effect of interaction between sample translations and the normalized alignment ratio (${\chi_{(245.9,276)}}^2 = 1983$, $p < .0001$) can also be found. The output is reported below in Table 10. As the output shows, both alignment ratio and the longer alignments have significant impacts on the fluency scores as well, and apparently three word BMWUs

Table 10: Fixed Effects of Alignment Ratio and Alignment Length on Fluency

Response Variable	Effect	Estimate	SD	Z	$Pr. > t$ or Z	AIC
Fluency	Intercept	5.15	0.74	6.98	< 0.0001	7149.1
	NrmALR	0.90	0.39	2.29	< 0.05	
	AL2	1.35	0.69	1.94	< 0.05	
	AL3	1.89	0.75	2.52	< 0.05	
	TrCorpS2	-0.05	0.09	-0.55	< 0.58	
	NrmALR:AL2	-0.09	0.47	-0.21	> 0.05	
	NrmALR:AL3	-1.49	0.75	-2.00	< 0.05	

have more weight over other two. The interaction of normalized BMWU alignment ratio and BMWU length also suggests that longer alignments contribute more to fluency than shorter alignments (one and two word alignments) in our case.

4.3. Implications

Generalized additive modelling suggests that BMWU alignments (two words and above) play a significant role in determining the quality of students' translations. There is also very strong indication that alignment length interacts with the normalized alignment ratios in students' translations and impact on their quality, i.e. the normalized alignment ratios of different lengths vary in their contribution to the quality scores for adequacy and fluency. As Table 1 shows, translators tend to resort to prefabricated translation pairs (e.g. BMWUs) available to them. This decision-making conforms to the idiom principle (Sinclair, 1991) or formulaic language (Wray, 2001), which help the translators produce native-like selections and reduce the cognitive processing effort.

However, it seems that our BMWU alignments have less effect on fluency. This contradicts our intuition but can be explained by the fact that alignment places more emphasis on correspondences, which are often oriented at semantic equivalence. In terms of the fluency scores, our BMWUs are relatively short (up to 4 words in this study), so longer units to capture the discourse markers, cohesion devices, etc.

5. Related Work

Recent years have seen attempts using word alignment information for translation quality estimation (QE), for either machine translation or human translation (Ueffing et al., 2003; Abdelsalam et al., 2016; Specia et al., 2015; Camargo de Souza et al., 2013; Bach et al., 2011; Popović et al., 2011; Popovic, 2012; Yuan et al., 2016) . As Abdelsalam et al. (2016) noted, the majority of these research focus on exploiting alignment related information for word-level QE. Among the few studies Abdelsalam et al. (2016), Camargo de Souza et al. (2013) and Yuan et al. (2016) actually try to tackle QE at the sentence-level or above, some features are too complex and not friendly interpretable to humans. For instance, Bach et al. (2011) use the source and target alignment context and even combine alignment context with PoS tags, and Camargo de Souza et al. (2013)

implement features, such as proportion of alignments connecting words with the same PoS tag and proportion of words in ST and TT that share the same PoS tag. We argue that on the one hand, such features are not computing cost-effective, and on the other hand, they are against our intuition that linguistic attributes, such as PoS, will hardly remain the same during the translation of two drastically different language pairs. We continue our way of obtaining alignment precision and recall in (Yuan et al., 2016), which compute the proportion of aligned words in source sentences or documents (precision) and the proportion of aligned words in target sentences or documents (recall), similar to two of many alignment features[6] by (Camargo de Souza et al., 2013). However, our BP alignment features differ by considering the sentence or document length information of ST and TT, and they are normalized [7]. Meanwhile, to be clear, we are not investigating how these features contribute in the QE task, but as part of feature engineering process, we explore statistically how different lengths and types of BP alignments interact with human rated translation quality scores (adequacy and fluency) to prove that they are useful in the future QE tasks.

6. Conclusion

In this study we investigated the effects of BMWUs extracted from parallel data to measure the adequacy and fluency of human translations.

Statistical analysis shows that the normalized alignment ratios for the phrase alignments longer than two words have the greatest impact on measuring translation quality. It is also found that longer phrase alignment ratios are statistically closer to adequacy than to fluency. We plan exploiting aligned BMWUs in the QE task (for evaluating both human and machine translations) in the form of normalized BMWU alignment ratios. These features can be used for QE at the sentence and document level. The latter task is particularly important, as the summative evaluation of trainee translators is typically done at the level of document translations, in contrast to conventional MT QE at the sentence level. Extending the alignment features to the phrasal level is consistent with human translation intuition and language production hypothesis, e.g. idiom principle (Sinclair, 1991). Most importantly, to the best of our knowledge, this paper is the first attempt of investigating the effects of phrasal alignment on human translation quality, and the method of computing alignment ratios has extended further beyond the word-level alignment features in previous MT QE studies. We share scripts from this study at `https://github.com/hittle2015/Bi-MWU.git`.

We can expect several extensions to the proposed model. First, our experiment uses general purpose parallel corpora, while the trainee translations come from a specific domain. It is interesting to investigate prediction of translation quality using parallel corpora from the same domain to measure the contribution of the proposed BMWU alignment ratios. Second, our experiment reported here accepts any phrase alignment from the professionally translated corpus as matching the trainee translations without taking their neighbouring contexts into consideration. We can try including a model for the context by using Recurrent Neural Networks methods from Neural Machine Translation when the neighbouring words contribute to the translation decisions (Koehn and Knowles, 2017).

Another extension for this study concerns increasing the amount of reliable BMWUs by extracting them from the comparable corpora (Sharoff et al., 2013), since the amount of data from monolingual corpora is much greater than what comes from parallel corpora, especially for specific domains. There has been extensive research on alignment of the monolingual embedding spaces for individual words, see an overview in (Conneau et al., 2017), but so far not much on BMWUs.

[6] They use proportion of aligned words and proportion of aligned n-grams. The latter is similar to our proposed feature of phrase alignment.

[7] We also propose that the summation of the logarithmized probabilities (IBM scores) of all aligned words in the documents (sentences) could be a potential quality indicator, and so is the geometric mean of these probabilistic scores.

7. Acknowledgements

This study is partially funded by the Jiangsu Provincial Social Science Fund (No.: 17YYB013) and the teaching reform project of Jinling Institute of Technology (No.: JYJG2017-34).

8. References

Abdelsalam, A., Bojar, O., and El-Beltagy, S. (2016). Bilingual embeddings and word alignments for translation quality estimation. In *Proceedings of the First Conference on Machine Translation*, pages 764–771, Berlin, Germany, August. Association for Computational Linguistics.

Abu-Ssaydeh, A.-F. (2004). Translation of english idioms into arabic. *Babel*, 50(2):114–131.

Bach, N., Huang, F., and Al-Onaizan, Y. (2011). Goodness: A method for measuring machine translation confidence. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Techologies*, pages 211–219, Portland, Oregon, USA, June. Association for Computational Linguistics.

Baker, M. (2011). *In other words: a coursebook on translation*. Routledge, London, 2nd edition.

Baobao, C., Danielsson, P., and Teubert, W. (2002). Extraction of translation unit from chinese-english parallel corpora. In *Proceedings of the First SIGHAN Workshop on Chinese Language Processing - Volume 18*, SIGHAN '02, pages 1–5, Stroudsburg, PA, USA. Association for Computational Linguistics.

Camargo de Souza, J. G., Buck, C., Turchi, M., and Negri, M. (2013). FBK-UEdin participation to the WMT13 quality estimation shared task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 352–358, Sofia, Bulgaria, August. Association for Computational Linguistics.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. *CoRR*, abs/1710.04087.

DeNero, J. and Klein, D. (2008). The complexity of phrase alignment problems. In *Proceedings of ACL-08: HLT, Short Papers*, pages 25–28, Columbus, Ohio, June. Association for Computational Linguistics.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia, June. Association for Computational Linguistics.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In *Proc Neural Machine Translation Workshop*, Vancouver.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Liu, I. and Agresti, A. (2005). The analysis of ordered categorical data: An overview and a survey of recent developments. *Test*, 14(1):1–73.

Min, F. (2007). Cultural issues in chinese idioms translation. *Perspectives*, 15(4):215–229.

Neubig, G., Watanabe, T., Sumita, E., Mori, S., and Kawahara, T. (2011). An unsupervised model for joint phrase alignment and extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 632–641, Portland, Oregon, USA, June. Association for Computational Linguistics.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Och, F. J., Tillmann, C., Ney, H., et al. (1999). Improved alignment models for statistical machine translation. In Pascale Fung et al., editors, *Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, College Park,MD,USA, 21-22 June.

Pitman, J. and Yor, M. (1997). The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900.

Popović, M., Vilar, D., Avramidis, E., and Burchardt, A. (2011). Evaluation without references: Ibm1 scores as evaluation metrics. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 99–103, Edinburgh, Scotland, July. Association for Computational Linguistics.

Popovic, M. (2012). Morpheme- and pos-based ibm1 and language model scores for translation quality estimation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 130–134, Montreal, Canada, June. Association for Computational Linguistics.

Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for nlp. *Computational Linguistics and Intelligent Text Processing*, pages 189–206.

Sharoff, S., Rapp, R., and Zweigenbaum, P. (2013). Overviewing important aspects of the last twenty years of research in comparable corpora. In Serge Sharoff, et al., editors, *BUCC: Building and Using Comparable Corpora*, pages 1–17. Springer.

Sinclair, J. M. (1991). *Corpus, concordance, collocation*. Oxford University Press, Oxford.

Specia, L., Paetzold, G., and Scarton, C. (2015). Multilevel translation quality prediction with quest++. In *ACL-IJCNLP 2015 System Demonstrations*, pages 115–120, Beijing, China.

Tian, L., Wong, D. F., Chao, L. S., Quaresma, P., Oliveira, F., and Yi, L. (2014). Um-corpus: A large english-chinese parallel corpus for statistical machine translation. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Ueffing, N., Macherey, K., and Ney, H. (2003). Confidence measures for statistical machine translation. In *Proceedings of the MT Summit IX*, New Orleans, USA, September. the Association for Machine Translation in the Americas.

Wang, W.-l. and Nian, X.-p. (2004). Loss of beauty in translation of idiom and compensation. *Hefei Gongye Daxue Xuebao Shehui Kexue Ban/Journal of Hefei University of Technology (Social Sciences)*, 18(6):118–120.

Wood, S. N., Pya, N., and Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111(516):1548–1563.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R, Second Edition*. Chapman & Hall / CRC texts in statistical science. Chapman and Hall/CRC, second edition edition.

Wray, A. (2001). *Formulaic language and the lexicon*. Cambridge University Press, Cambridge.

Yuan, Y., Sharoff, S., and Babych, B. (2016). Mobil: A hybrid feature set for automatic human translation quality assessment. In Nicoletta Calzolari, et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, May. European Language Resources Association (ELRA).

Zebardast, M. R. and AbuSaeedi, A. A. R. (2015). Cultural loss versus hermeneutic of translation: A case study on the idioms of "death of a salesman" (miller, 1949). *Journal of Academic and Applied Studies*, 5(11).

Zebardast, M. R. (2015). Perceptions about cultural loss in translating idioms from english into persian: A case study on the “death of a salesman” (miller, 1949). *Journal of Academic and Applied Studies*, 5(10).