A Framework to Build Quality into Non-Expert Translations

Christopher G. Harris[1]

[1] Dept. of Computer Science, University of Northern Colorado, Greeley CO 80639 USA
christopher.harris@unco.edu

Abstract. When performed correctly, crowdsourcing can produce near-expert quality translations both quickly and inexpensively; however, the quality obtained from crowdworkers is rarely consistent. We propose a framework to obtain high-quality work from non-expert crowdworkers by incorporating intermediate mechanisms such as ranking and editing in addition to translation. We conduct three empirical experiments in which we explore the impact of these framework mechanisms on translation quality, time, and cost. We also demonstrate that our proposed framework is robust against spammers, verifiable at different steps, and consistent with little variance in quality. Our framework achieved a higher BLEU score than professional translators at a fourteenth of the cost but required about 50 percent more time to complete. Therefore, it is most appropriate when a task requester wishes to maximize translation quality and minimize cost.

Keywords: translation quality, crowdsourcing, translation framework

1 Introduction

Machine translation (MT) tools like Google Translate have recently made impressive strides in quality by implementing deep learning techniques. However, MT tools are unable to match the nuanced advantages that professional translations can provide, such as avoiding mistranslations, applying political and social correctness to translated text, and for many free online tools, maintaining confidentiality. One issue with professional translations is that the costs can be excessive, particularly with low-resource languages. To translate a corpus from Tamil to English, Germann (2001) calculated the cost of each translated word at $0.36. However, a more recent estimate of the cost of translation for more common languages such as Spanish, French, and Chinese (obtained from websites proz.com, slator.com, and wordminds.com) range from $0.27 to $0.33 per word. These resources demonstrate that professional translation tasks between common languages can quickly become prohibitively expensive, even for moderate-sized translation efforts. Automatic evaluation metrics such as BLEU (Papineni et al., 2002) have been shown to correlate well with human evaluations for machine translations (e.g., Hobson et al.,

2007, Wu et al., 2016). Unfortunately, payments made for non-expert translations are often weakly correlated with the resulting output quality.

The growth of the internet has led to a noteworthy increase of choices for translations – a substantial number of non-experts, especially crowdworkers, were now available to perform the same tasks previously relegated to experts. Crowdsourcing platforms such as Amazon Mechanical Turk (AMT) have become prominent marketplaces for translations, particularly because they offer easy matchmaking services between task requester and worker. These platforms promote the benefits of having many tasks and workers available at any one time, offering inexpensive labor to requesters (and new opportunities to earn money for workers). These platforms focus on smaller, self-contained tasks with simple instructions (called microtasks) that can be performed quickly. It is often up to the requesters to provide checks on quality, which can be challenging when translating from a language the requester knows to one that they may not understand (or vice versa).

An ongoing concern regarding the use of crowdsourcing platforms such as AMT is that the quality of outputs can vary considerably. Moreover, there is an inherent misalignment between task requester and crowdworker; while both task requesters and workers desire the task to be accomplished quickly, most task requesters seek to maximize quality and minimize cost. However, since crowdworkers are perceived to be anonymous and transient, a sizeable subset of them, called spammers, seek to maximize their earnings with little concern for the output quality, giving rise to a market of imperfect information (Akelof, 1978). From a requester's perspective, mechanisms to maximize quality and speed while minimizing cost need to be designed into any robust text translation system. Few requesters know how to do this or understand which mechanisms will provide the largest improvement in translation quality.

Over the last decade, researchers have tried a variety of approaches to maximize output quality while simultaneously minimizing cost. Although each researcher has independently demonstrated techniques to provide near-expert translation quality, there has been a scant focus on providing a comprehensive framework or set of tools to enhance these efforts. Moreover, none of these approaches examined the optimal situation in which to employ each mechanism. In this paper, we seek to provide these, as well as to answer the following research questions.

1. Which components of our proposed framework can provide the greatest increase in translation quality with respect to time and cost?

2. How does adding additional crowdworkers in each step of the framework affect quality, time and cost?

Our paper makes the following contributions. We define distinct components of crowd-based translation tasks and examine how the flow of these components in a framework can affect quality. We examine these framework components in turn and discuss whether the use of each is most appropriate based on time, cost, and improvement in quality. We conduct experiments using non-experts (crowdworkers) that empirically examine each metric in our framework. We discuss our findings with respect to the overall framework and crowdsourcing in general.

2 Background and Related Work

For over a decade, researchers have shown that non-experts hired from crowdsourcing platforms have been able to translate text with quality that approaches those produced by experts. Kittur et al. (2008) were one of the first to compare crowdworkers to experts in an NLP task. Workers were asked to rate Wikipedia articles according to several factors, such as the quality of writing, accuracy, and structure, on a seven-point Likert scale. Initially crowdworker ratings were poorly correlated with those made by experts (r=0.5); however, by redesigning their experiment to improve quality, the correlation with the expert ratings improved in a subsequent experiment (r=0.66). They suggested designing tasks to encourage accurate responses over malicious or random ones (i.e., spam), but they did not discuss how task flow can improve quality.

Snow et al. (2008) used AMT for non-expert annotations for five NLP tasks. Their majority voting approach only required a few non-expert labels to achieve near-expert quality. Callison-Burch (2009) used AMT to evaluate MT outputs using a few targeted strategies. In one approach, they hired non-experts to create reference translations in several languages, reinforcing the ease of obtaining near-expert quality translations quickly and inexpensively.

Cost savings using crowdsourcing platforms can be substantial, even when redundant quality checks are designed into the system. Hoffmann (2009) indicates that using crowdworkers for translating transcription services can save a company 33% compared with using in-house staff. Harris and Xu (2011) found using non-expert translators for translated transcriptions from Chinese to three other languages were, on average, 1/23rd the cost of professional translators, but translation quality was comparable. Novotney and Callison-Burch (2010) found professional translation costs to be thirty times as expensive as using crowdworkers.

The relationship between compensation and quality is often confounding. For example, Gillick and Liu (2010) experimented with different amounts of compensation for a text summarization exercise and found that a lower compensation ($0.07) resulted in better quality. They surmised that a lower amount attracted crowdworkers who prioritized quality over making money,

although the pool size of crowdworkers with such a priority on quality was small. However, in another study, Aker et al. (2012), found that higher payments resulted in better quality for objective tasks that contained verifiable answers (e.g., performing a calculation). In our study, we examine the relationship between quality and compensation provided to non-expert translators.

Using crowdworkers to do portions of a task has been examined previously, but only as inputs to MT algorithms. Buzek et al. (2010) used AMT to create paraphrase lattices as MT inputs. Two tasks were established: one to create the paraphrase lattices, and another one to verify the generated paraphrases. They found that if paraphrasing by crowdworkers targeted the most challenging areas of the lattice, TER score of the resulting translation could be improved.

A few researchers have explored the use of the crowd for different subtasks of a single translation project. Bloodgood and Callison Burch (2010) produced test sets for MT systems using crowdworkers instead of professionals and found that the quality of these crowd-created inputs matched those made by professionals. Zaidan and Callison-Burch (2009) examined the benefits of redundancy in translating, editing, and voting. They concluded that redundancy provides tangible benefits, but they do not indicate the relative benefits. Yan et al. 2016 examined creating a two-stage model with translator-editor pairs on AMT. They found that randomly pairing translators and editors provided the best quality. Hourcade and Gehrt, (2015) used crowdworkers in a two-step process: first to summarize medication warnings and then vote on the best summarization. El-Haj et al. (2017) obtained semantic labels for 250 words in six languages finding that a two-stage filtering process they used with crowdworkers improved quality and reduced spam. We build upon this notion in this paper.

The common limitation in these previous studies is that they focus on the use of crowdworkers only with specific components of translation and text summarization tasks – and do so convincingly – but do not provide an overall framework or guidance to follow. Our goal is to derive a framework that can be followed to build quality into translations from start to finish.

3 Crowdsourcing Framework

Translations frequently involve multi-lingual and bi-lingual workers, each of whom translates a snippet of text in a source language into snippets of text in (one or more) target languages. Text summarization, on the other hand, only requires monolingual non-experts, thus appealing to a larger pool of crowdworkers. MT tools such as Google Translate are improving, but still lack the nuances of a human translator; to this end, we focus on providing translations with a focus on quality. Since crowdsourcing platforms are designed to provide microtasks or simple tasks each requiring a few minutes at a time, we also focus on keeping each task as atomic as possible with simple instructions. We seek to develop a framework with the following qualities:

- **Robust:** Our framework should be impervious to low-quality inputs from a malicious crowdworker or spammer.
- **Verifiable:** We should be able to evaluate our metrics after each crowdworker-dependent step in our framework.
- **Consistent:** To the extent possible, the task should be deterministic; the same inputs should produce approximately the same outputs, even with a different set of crowdworkers.

3.1 Framework Components

Five crowdsourcing-dependent components are used in our proposed framework. To keep the pool of potential crowdworkers as large as possible for translations (which, due to the economics of supply and demand, lowers the cost and reduces the task completion time), we want as few components as possible to rely exclusively on multi- and bi-lingual crowdworkers

- **Ranking:** This component asks crowdworkers to rank text in order of relative preference. Ranking is helpful in situations where users have few choices and can discriminate between the choices; if there are many choices to choose from, scoring each item (on a Likert scale) may make more sense than ranking them. Studies using crowd ranking, such as that by Goto (2015) have demonstrated their effectiveness. Ranking does not depend on multi- or bi-lingual crowdworkers.
- **Translation:** Considered the core component, this is the task in which translated versions of the input text are generated.
- **Editing:** When a translation is achieved by using different crowdworkers, the overall tone and flow are affected. Editing "smooths" the document by improving the flow between segments of text, removing redundancies, and making the tone of the article consistent. Editing does not depend on multi- or bi-lingual crowdworkers.
- **Disassembly:** Divide a document (or collection of documents) into segments (subsets of the document of approximately consistent size). Disassembly is usually accomplished through automation.
- **Reassembly:** Recombine the translated segments into a single document. Like disassembly, reassembly is usually accomplished through automation.

3.2 Framework Flow

We begin with the document in step 1 of Figure 1. For discussion, we consider it a single document although it could be comprised of a set of documents.

In step 2, the document is automatically broken into *n* segments (t_1, t_2, … t_n) – each segment can be as small as a sentence or as large as a paragraph. In step 3, each segment is submitted separately for translation by *m* non-experts. We note that for smaller segment sizes (i.e., a sentence), the lack of context may make translations challenging; therefore, we provide the entire document to each translator. This results in *m* translations for each segment, (e.g., $t_{n(1)}$, $t_{n(2)}$, … $t_{n(m)}$ would be the translated segments for t_n).

Next, for each of the *n* segments, the *m* translated segments are then ranked by a separate pool of crowdworkers (step 4). The use of a separate pool of crowdworkers helps insure that the output of a malicious crowdworker or spammer does not propagate past this stage. It is unique to our framework.

Once the *m* translations are ranked 1…*m*, the highest-ranked segment for each of the *n* segments (denoted as $t_{1(1)}$' … $t_{n(1)}$' or simply t_1'… t_n') is retained. These are then automatically reassembled in the original order (step 5). Although at this point we have a translation of the full document, it is really a rough combination of translations of different segments of text and is unlikely to flow together well. To accomplish this, we edit the translated document to ensure the translation maintains the correct context and follows a consistent tone. We accomplish this by having the document edited by *q* crowdworkers (step 6), and these outputs are ranked by another pool of crowdworkers (step 7) to obtain the final document (step 8). This final document is evaluated for quality using BLEU. Figure 2 provides an example showing a Chinese document divided into segments of a single sentence each.

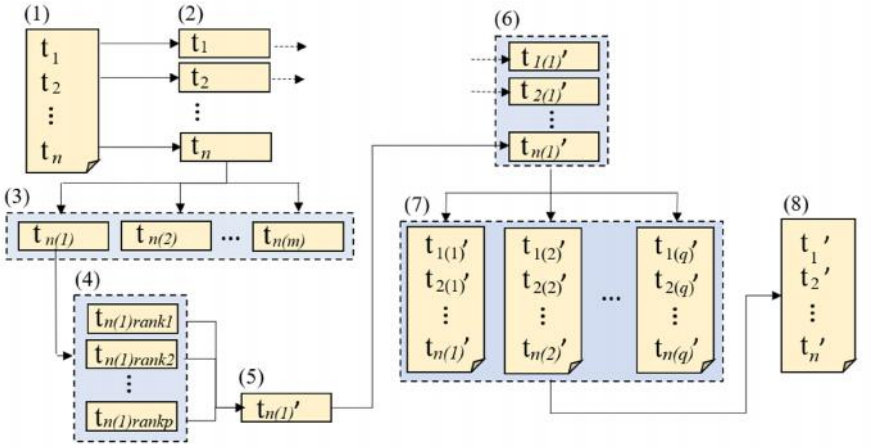

Fig. 1. Our framework for conducting translations and text summarizations using non-experts. Key tasks include translations (step 3), ranking translations (step 4), reassembly of the top-ranked translations for each segment (step 6) and a ranking of edits of the reassembled segments (step 7).

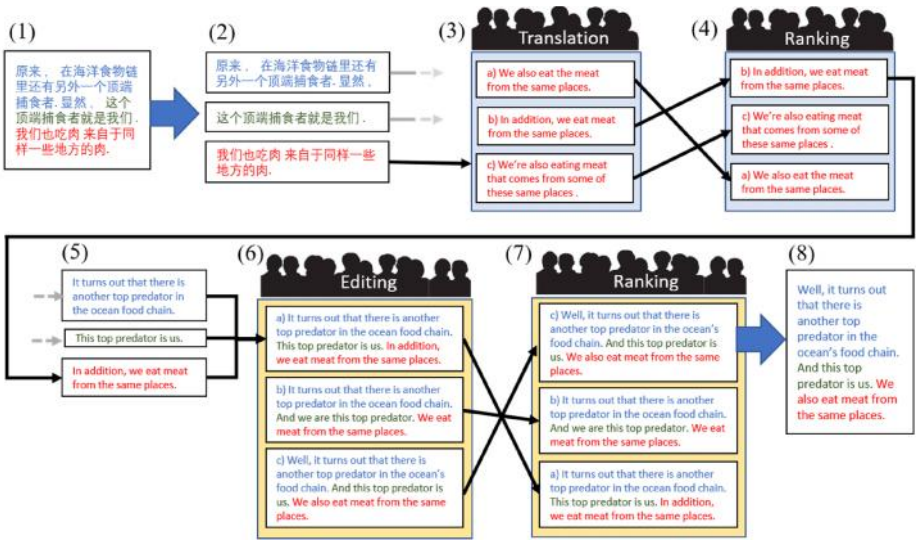

Fig. 2. An example of the 8 steps of the crowdsourcing framework given in Figure 1.

4 Experiment Design

With many variables to consider, we wish to optimize our framework. We conduct experiments to determine the payment amount, the number of crowdworkers, and the segment size necessary to produce expert-quality translations.

4.1 Data

We use a publicly-available dataset: the TED 2013 parallel corpora from (Tiedemann, 2012). We randomly select three transcripts, containing an average of 124 sentence pairs and 2299 English words each. For these transcripts, we obtain the parallel corpus for Chinese-English and French-English. As in Callison-Burch (2009), our objective was to reduce the use of available MT tools by crowdworkers, which would be viewed as cheating; we provided images of each text segment to prevent workers from copying and pasting text into an MT tool.

4.2 Metrics

We use BLEU to compare translations against the gold standard. Other studies on MT use other metrics such as inter-annotator agreement, but this would be difficult to implement in our case. We also record the average time taken in increments of 10 minutes.

By design, the framework components involving the crowd were required to be done consecutively. Some experiments involved examining tradeoffs between quality and payment amount; running these experiments at the same time could introduce bias. To counter this, we staggered the start times of the crowd-based tasks when different values or amounts were being considered.

We devised a process to automatically list ranking tasks on AMT once all segments were translated and when edits were done. We started the clock when a task was listed on AMT and end when all tasks for that step were completed. We sum the time taken for all tasks. The automated steps of disassembly (step 2) and reassembly (step 5) are assumed to be done instantaneously.

To determine the highest-ranked segment by the crowd in the ranking steps (steps 4 and 7), we use a Borda count (Saari, 1999). Borda counts take a rank of p candidate segments; each candidate ranked highest will receive p points, those ranked second will receive $p-1$ points, etc. These counts are summed across all r ranked lists to obtain a total count, with the highest total score selected as the best translation. When p and r are small, Borda counts may lead to ties, so a plurality vote (the largest number of first-place votes) and anti-plurality votes (i.e., the smallest number of last-place votes), in that order, serve as tiebreakers. Indeed, this method resolved all ties in rankings made by the crowd.

4.3 Gold Standard and Baselines

The translation for each transcript provided in the parallel corpora is used as our gold standard. Also, we asked two professional translators to translate each transcript to English at the average market translation rate of $0.30/word. We asked the translators to provide the time it took to translate each transcript, and we average the times they provided in our analysis. To translate the three transcripts from the two languages, the total cost for professional translation is $4137.60, representing an average hourly rate of $36.94. We use this as our first baseline.

In many translation studies involving non-experts, a single worker is asked to translate an entire document or transcript. We wish to see how this larger task would compare to our framework's use of microtasks. Crowdworkers were asked to translate documents from French and Chinese to English. We hired three crowdworkers for each language pair and asked each to translate a

single transcript for $40.00 per transcript ($0.017/word). We used the Upwork website, which provides task requesters a method to track the actual time workers spent working on their tasks. We paid a total translation cost of $240.00 (excluding the Upwork platform fees), which calculates to an average wage of $5.46/hr. We use this as our second baseline.

4.4 Participants

Participants were hired from AMT. They were required to have an overall approval rate of 90%. Translators were shown the entire transcript in either French or Chinese; the text segment they were asked to translate was marked. Editors were shown the entire document with each segment identified and asked to make the combined document flow smoothly, paying careful attention to tone and flow. Rankers were shown between 3 and 12 segments (depending on the experimental conditions) and asked to rank the segments from best to worst. If a crowdworker did not complete the task according to the rules, we removed the results and relisted the task. We had only a single case where the task had to be repeated.

5 Experiments

5.1 Experiment 1: The Effects of Payment Amount

On average, each transcript is divided into 12 segments of 192 words each. We wished to see if the amount we offered affected the quality and time. Using the three transcripts, we varied the payment amounts presented to crowdworkers offering $0.25, $0.50, and $1.00 to translate one segment (step 3). We offered another set of crowdworkers the same amounts to edit documents for context and flow (step 6). For the steps involving ranking, we paid a consistent $0.25 each and had three crowdworkers rank each item (in steps 4 and 7). We evaluate the documents at two different points, just after the document is reassembled with each highest-ranked segment in step 6 (but before editing), and at when the final translation has been completed at step 8. Table 1 provides an evaluation of each payment amount on the average time and quality.

	Amt Paid Per Trans Doc	**French to English**				**Chinese to English**			
		Pre-editing (step 6)		**Post-editing (step 8)**		**Pre-editing (step 6)**		**Post-editing (step 8)**	
		BLEU	**Time**	**BLEU**	**Time**	**BLEU**	**Time**	**BLEU**	**Time**
Part 1: Varying the Pmt Amts for Translation and Editing, but keeping Ranking pmts @$0.25 ea									
Trans/Edit @$0.25	$36.00	22.17	11:40	37.55	24:20	21.75	11:40	36.06	26:20
Trans/Edit @$0.50	$54.00	**24.05**	8:50	**38.04**	19:30	22.89	9:30	36.39	20:20
Trans/Edit @$1.00	$90.00	23.97	**7:10**	37.78	**17:50**	**23.06**	7:50	**36.44**	**19:10**
Part 2: Varying the Pmt Amts for Ranking, but keeping Translation and Editing pmts @$0.50 ea									
Ranking @$0.10	$43.20	23.03	10:20	37.84	21:40	22.15	9:50	35.91	22:50
Ranking @$0.25	$54.00	**24.05**	8:50	**38.04**	19:30	22.89	9:30	36.39	20:20
Ranking @$0.50	$72.00	23.52	**7:40**	37.93	**19:00**	**23.01**	**10:50**	**36.43**	**19:40**

Table 1. BLEU scores and time taken (in hours) for the pre-editing and post-editing step in our framework, broken out by language.

From Part 1 of Table 1, there is only a marginal difference in quality (as determined by BLEU score) regardless of the amount we pay the crowdworkers for translation and editing. This backs up the findings of Mason and Watts (2010) on crowdworker payments for other types of tasks. We also notice that there is a jump in the BLEU score between the pre-editing step and the post-editing step irrespective of the amount paid, demonstrating the merits of the editing step in our framework. From a task requester's perspective, offering $0.50 for translation and editing seems to be an appropriate tradeoff between payment and quality; we carry this forward into our following experiments.

The amount of payment offered does affect the task completion time. This makes sense; some crowdworkers have minimum amounts they will accept before engaging in a task. Completion times for Chinese-to-English translations took more time than those from French to English; this is likely an artifact of the relatively smaller number of Chinese translators available on AMT.

In Part 2 of this experiment, we hold payments for translation of each segment and editing of each document constant at $0.50 and evaluate the effects of the payment amount for ranking has on quality and time. For each transcript, we offer crowdworkers the payment amounts of $0.10, $0.25, and $0.50 to rank (steps 5 and 7). The results of ranking payment amounts on the average time, and quality for each language are provided in Part 2 of Table 1.

The BLEU score does not change much in the ranking steps even when we offer five times as much compensation. Although the time taken decreases when we pay more for ratings, the reduction in time is not as sensitive to cost as it is for editing and translations. We keep payments for rankings at $0.25 for follow-on experiments as this seems to provide the best balance between quality and payment.

	Amt Paid Per Trans Doc	**French to English**				**Chinese to English**			
		Pre-editing (step 6)		**Post-editing (step 8)**		**Pre-editing (step 6)**		**Post-editing (step 8)**	
		BLEU	**Time**	**BLEU**	**Time**	**BLEU**	**Time**	**BLEU**	**Time**
Varying the Size and Number of Segments Translated									
2 segments, avg. 1149 words ea.	$24.00	23.62	**5:20**	37.28	**14:40**	22.49	**5:40**	36.11	**18:00**
4 segments, avg. 575 words ea.	$30.00	23.77	6:50	38.01	16:30	22.73	6:40	36.43	18:10
8 segments, avg. 287 words ea.	$42.00	23.76	8:00	**38.12**	18:10	22.81	8:40	**36.56**	19:30
12 segments, avg. 192 words ea.	$54.00	**24.05**	8:50	38.04	19:30	**22.89**	9:30	36.39	20:20

Table 2. The effects of varying the size and number of segments on BLEU scores and time taken (in hours) for the pre-editing and post-editing step in our framework, broken out by language.

5.2 Experiment 2: The Effects of Segment Size

Larger segment sizes provide a single translator with more context to work with, which is likely to improve translation quality. On the other hand, smaller segment sizes are more suitable to crowdsourcing platforms, which are designed for *microtasks*, or small tasks that can be accomplished quickly and inexpensively. To examine the effects of segment size, we divide up each of the three transcripts into segment sizes of approximately 1000 words, 500 words, and 250 words each, representing 2, 4 and eight segments per transcript. We increase the payments for translation (step 3) to correspond with the increase in task size (recall we paid $0.50 for each of 12 translated segments of 192 words in Experiment 1); corresponding payments are $3.00, $1.50, and $0.75 per segment. We pay a constant $0.50 for editing the translated text and $0.25 for ranking. The results are shown in Table 2.

From Table 2, using larger segment sizes decreases the time taken, but overall it has little effect on the quality of the resulting translation. This seemed puzzling at first glance, but it shows the value of the framework's editing and ranking steps. We also note that having more segments reduced the variance in quality as well, although this is not shown in the table. We notice that the best quality is achieved with eight segments and we carry this forward in our next experiment.

5.3 Experiment 3: The Effects of the Number of Crowdworkers

We have used three crowdworkers at each step in the previous experiments – would increasing or decreasing the number of crowdworkers used in each step influence quality and time taken? We vary the number of crowdworkers as shown in Table 3.

			Amt Paid Per Trans Doc	French to English				Chinese to English			
				Pre-editing (step 6)		Post-editing (step 8)		Pre-editing (step 6)		Post-editing (step 8)	
				BLEU	Time	BLEU	Time	BLEU	Time	BLEU	Time
Varying the Number of Translators, Editors and Rankers											
Trans	Ed	Rank									
3	1	1	$26.00	21.77	**7:40**	36.64	**17:10**	21.30	**8:10**	32.92	**17:50**
3	3	1	$34.00	21.77	**7:40**	37.79	17:40	21.30	**8:10**	36.09	19:20
3	3	3	$42.00	**24.05**	8:50	38.04	19:30	22.89	9:30	36.39	20:20
3	5	3	$50.00	**24.05**	8:50	**39.45**	20:30	22.89	9:30	**37.11**	20:40
5	5	3	$54.00	23.98	10:10	**39.45**	22:00	**24.59**	10:20	**37.11**	21:30
5	5	5	$58.00	23.98	10:30	**39.45**	22:20	**24.59**	10:40	**37.11**	21:50

Table 3. The effects of varying the number of crowdworkers for each task on BLEU scores and time taken (in hours) for the pre-editing and post-editing step in our framework task, broken out by language.

Quality is most sensitive to the number of editors and least sensitive to the number of rankers. We notice there is no benefit to increasing the number of rankers from three to five. There is also no benefit to increasing the number of translators beyond three. The translation step is the most sensitive to the amount of time taken, is the bottleneck in our framework, and is the only step dependent on bi- and multi-lingual crowdworkers; adding additional translators slows the overall translation process down.

6 Analysis

Using a framework design with three translators, five editors, and three rankers, which provides a balance between payment amount and quality, we compare our results with those obtained by the other baselines. Table 4 shows this comparison.

	Amt Paid Per Trans Doc	**French to English**			**Chinese to English**		
		Avg BLEU	**SD BLEU**	**Avg Time**	**Avg BLEU**	**SD BLEU**	**Avg Time**
Professional Translator Baseline	$689.60	38.23	7.09	13:10	36.70	5.63	13:30
Single Crowdworker Baseline	$40.00	33.35	8.53	**17:00**	31.84	7.73	**17.40**
Framework	$50.00	**39.45**	**3.56**	20:30	**37.11**	**3.28**	20:40

Table 4. A comparison of our framework's results with our two baselines. We compare the average and standard deviation (SD) of BLEU score, and the average time taken, by language.

Our framework achieved a BLEU score that exceeds the professional translations for both languages to English at a fourteenth of the cost. However, our framework approach took nearly 1.5 times as long to complete as the quicker of the two baseline approaches. Our framework also achieved a BLEU score that was convincingly better than the single crowdworker baseline, but as with the professional translator, this comes with the tradeoff of requiring much more time. Additionally, the translation cost using our framework was triple that of the single crowdworker; with quality as our framework's paramount consideration, paying more for better quality is an appropriate tradeoff. Our second experiment had shown that, even when our framework used larger segment sizes (approximating the single crowdworker), it was superior in quality to the single crowdworker based on the BLEU scores. This demonstrates the value of a divide-and-conquer approach used by the framework and how it helps low-quality inputs from malicious crowdworkers and spammers from affecting quality.

One often-stated benefit of crowdwork is that tasks can be done quickly with sufficient quality and at a low price. For the metrics of cost, quality and time, our experiments show it is a challenge to achieve all three simultaneously. Our framework is best designed for those with a focus on cost and quality

at the expense of time. The single crowdworker model would cost less to implement and take far less time, but our experiments show that quality is significantly inferior. The professional translator would work best when quality and time are important, but the cost is not an issue.

We have been able to examine the time and quality at different points of our framework, demonstrating the verifiability principle. The standard deviation for the crowdworker baseline approach is more than double that of our framework, demonstrating the framework's robustness. It also means our results are not dependent on only a small subset of crowdworkers delivering quality outputs and obscuring the spam from the remaining crowdworkers, which is a common problem with crowdwork.

7 Conclusion and Future Work

We have introduced a framework for implementing expert quality for translations that approach professional quality at a lower cost of hiring a professional. This framework extends the work of many previous researchers but adds checks for robustness, verifiability, and consistency. Proponents of crowdwork often state the benefits are quality work that is cheaper and quicker than work provided by experts. Our experiments have shown that only two of the three are achievable; using a single crowdworker provides lower quality, a professional translator is higher cost, and our framework takes significantly more time. Expert quality can be achieved using our framework and as few as three translators and five document editors for editing, along with three other crowdworkers to rank the outputs from each of these steps. Using this configuration, we obtained higher quality than a professional at about a sixth of the cost. The tradeoff is that the configuration required more time than a professional translator or a single translator hired from a crowdsourcing platform.

In future work, we plan to examine how low resource languages might affect quality, cost and time taken by crowdworkers. Our work focused on using AMT as our crowdsourcing platform, but there are many other platforms now available which support entirely different demographics. Evaluating a variety of platforms will allow us to examine the robustness principle in greater detail. We also plan to examine how to better integrate MT tools in our framework to determine if we can improve the quality at an even lower cost. We also plan to examine hybrid approaches like those proposed by Borromeo et al. (2016). These authors performed some initial translations using Google Translate and then asked crowdworkers to make edits to those translations. This could be easily incorporated into our framework. We expect this approach to lower costs and decrease completion time while maintaining high quality in our framework as well. We also plan on examining the work of Gao et al. (2015). They have devised a method to stop once they receive a

translation that meets a minimum score threshold. We could incorporate this into our proposed framework, and we expect this would also lower costs.

It may be surprising to note that the professional translators did not achieve better BLEU scores than the framework we introduced. This may be an artifact of the source of the TED 2013 corpora, which used volunteer transcriptions and translations from the TED web site. To ensure the translation quality was at the professional level, we had translators unaffiliated with this study examine the translations, and they verified they were of professional quality. This may call into suspicion the BLEU metric as a fair assessment tool, which is beyond the scope of this study.

Our framework could also apply to other NLP tasks such as text summarization. We are currently conducting experiments to examine if our framework can ensure robustness, verifiability, and consistency in text summarizations just as it does for translations. Initial findings are positive.

References

1. Ahmet Aker, Mahmoud El-Haj, M-Dyaa Albakour, and Udo Kruschwitz. Assessing Crowdsourcing Quality through Objective Tasks. In *8th Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey 2012.
2. George A. Akerlof. "The market for “lemons”: Quality uncertainty and the market mechanism." In *Uncertainty in Economics*, pp. 235-251. 1978.
3. Michael Bloodgood and Chris Callison-Burch. "Using Mechanical Turk to build machine translation evaluation sets." In P*roceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 208-211. ACL 2010.
4. Ria Mae Borromeo, Maha Alsayasneh, Sihem Amer-Yahia, and Vincent Leroy. Hybrid Deployment Strategies for Crowdsourced Text Creation Tasks. In *Technical Report, University of Grenoble-Alps*, 2016.
5. Olivia Buzek, Philip Resnik, and Benjamin B. Bederson. "Error driven paraphrase annotation using mechanical turk." In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 217-221. ACL, 2010.
6. Chris Callison-Burch "Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk." In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*, pp. 286-295. ACL, 2009.
7. Mahmoud El-Haj, Paul Rayson, Scott Piao, and Stephen Wattam. "Creating and validating multilingual semantic representations for six languages: expert versus non-expert crowds." In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. ACL, p. 61-71. Valencia, Spain 2017
8. Ulrich Germann, Michael Jahr, Kevin Knight, Daniel Marcu, and Kenji Yamada. "Fast Decoding and optimal decoding for machine translation." In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pp. 228-235. ACL, 2001
9. Dan Gillick and Yang Liu. "Non-expert evaluation of summarization systems is risky." In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 148-151. ACL, 2010.
10. Mingkun Gao, Wei Xu, and Chris Callison-Burch. "Cost optimization in crowdsourcing translation: Low-cost translations made even cheaper." In *Proceedings of the 2015 Confer-*

ence of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 705-713. 2015.

11. Shinsuke Goto, Donghui Lin, and Toru Ishida. "Crowdsourcing for Evaluating Machine Translation Quality." In *10th Language Resources and Evaluation (LREC 2014)*, pp. 3456-3463. 2014.
12. Christopher G. Harris and Tao Xu. 2011. The importance of visual context clues in multimedia translation. In *International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF 2011)* (pp. 107-118). Springer Berlin Heidelberg.
13. Leah Hoffmann. 2009. Crowd control. Communications of the ACM 52(3) pp 16-17. DOI=http://dx.doi.org/10.1145/1467247.1467254
14. Juan Pablo Hourcade and Laura Gehrt. "Crowdsourcing for delivering research results to patients." In *Proceedings of HCI Korea*, pp. 196-202. Hanbit Media, Inc., 2014.
15. Aniket Kittur, Ed H. Chi, and Bongwon Suh. "Crowdsourcing user studies with Mechanical Turk." In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 453-456. ACM, 2008.
16. Winter Mason, and Duncan J. Watts. "Financial incentives and the performance of crowds." *ACM SigKDD Explorations Newsletter* 11, no. 2 (2010): 100-108.
17. Scott Novotney and Chris Callison-Burch. "Shared task: crowdsourced accessibility elicitation of Wikipedia articles." *In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. ACL, 2010.
18. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "BLEU: a method for automatic evaluation of machine translation." *In Proceedings of Computational linguistics*, pp. 311-318. Association for Computational Linguistics, 2002.
19. Donald G. Saari "Explaining all three-alternative voting outcomes." *Journal of Economic Theory*, 87, no. 2 (1999): 313-355.
20. Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. "Cheap and fast---but is it good? evaluating non-expert annotations for natural language tasks." In *Proceedings of empirical methods in natural language processing*, pp. 254-263. ACL, 2008.
21. May Al Sohibani, Najd Al Osaimi, Reem Al Ehaidib, Sarah Al Muhanna, and Ajantha Dahanayake. "Factors that influence the quality of crowdsourcing." In *New Trends in Database and Information Systems II*, pp. 287-300. Springer, Cham, 2015.
22. Jörg Tiedemann, 2012, Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*
23. Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, et al. "Google's neural machine translation system: Bridging the gap between human and machine translation." *arXiv preprint* arXiv:1609.08144 (2016).
24. Rui Yan, Mingkun Gao, Ellie Pavlick, and Chris Callison-Burch. "Are Two Heads Better than One? Crowdsourced Translation via a Two-Step Collaboration of Non-Professional Translators and Editors." In *ACL* (1), pp. 1134-1144. 2014.
25. Omar F. Zaidan and Chris Callison-Burch. "Crowdsourcing translation: Professional quality from non-professionals." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 1220-1229. ACL, 2011.