The Value of Monolingual Crowdsourcing in a Real-World Translation Scenario: Simulation using Haitian Creole Emergency SMS Messages

Chang Hu[‡], Philip Resnik[†‡], Yakov Kronrod[†]
Vladimir Eidelman[‡], Olivia Buzek[†‡], Benjamin B. Bederson[‡]
[†]UMIACS and Department of Linguistics
[‡]UMIACS and Department of Computer Science
University of Maryland, College Park
{changhu,bederson}@cs.umd.edu
{resnik,vlad,buzek}@umiacs.umd.edu
yakov@umd.edu

Abstract

MonoTrans2 is a translation system that combines machine translation (MT) with human computation using two *crowds* of monolingual source (Haitian Creole) and target (English) speakers. We report on its use in the WMT 2011 Haitian Creole to English translation task, showing that MonoTrans2 translated 38% of the sentences well compared to Google Translate's 25%.

1 Introduction

One of the most remarkable success stories to come out of the January 2010 earthquake in Haiti involved translation (Munro, 2010). While other forms of emergency response and communication channels were failing, text messages were still getting through, so a number of people came together to create a free phone number for emergency text messages, which allowed earthquake victims to report those who were trapped or in need of medical attention. The problem, of course, was that most people were texting in Haitian Creole (Kreyol), a language not many of the emergency responders understood, and few, if any, professional translators were available. The availability of usable translations literally became a matter of life and death.

In response to this need, Stanford University graduate student Rob Munro coordinated the rapid creation of a crowdsourcing framework, which allowed volunteers – including, for example, Haitian expatriates and French speakers – to translate messages, providing responders with usable information in as little as ten minutes. Translations may not have been perfect, but to a woman in labor, it had to have made a big difference for English-speaking responders to see *Undergoing children delivery Delmas 31* instead of *Fanm gen tranche pou fè yon pitit nan Delmas 31*.

What about a scenario, though, in which even amateur bilingual volunteers are hard to find, or too few in number? What about a scenario, e.g. the March 2011 earthquake and tsunami in Japan, in which there are many people worldwide who wish to help but are not fluent in both the source and target languages?

For the last few years, we have been exploring the idea of *monolingual* crowdsourcing for translation – that is, technology-assisted collaborative translation involving crowds of participants who know only the source or target language (Buzek et al., 2010; Hu, 2009; Hu et al., 2010; Hu et al., 2011; Resnik et al., 2010). Our MonoTrans2 framework has previously shown very promising results on children's books: on a test set where Google Translate produced correct translations for only 10% of the input sentences, monolingual German and Spanish speakers using our framework produced translations that were fully correct (as judged by two independent bilinguals) nearly 70% of the time (Hu et al., 2011).

We used the same framework in the WMT 2011 Haitian-English translation task. For this experiment, we hired Haitian Creole speakers located in Haiti, and recruited English speakers located in the U.S., to serve as the monolingual crowds.

2 System

MonoTrans2 is a translation system that combines machine translation (MT) with human computation (Quinn et al., 2011) using two "crowds" of monolingual source (Haitian Creole) and target (English)

Proceedings of the 6th Workshop on Statistical Machine Translation, pages 399–404,
Edinburgh, Scotland, UK, July 30–31, 2011.

speakers.[1] We summarize its operation here; see Hu et al. (2011) for details.

The Haitian Creole sentence is first automatically translated into English and presented to the English speakers. The English speakers then can take any of the following actions for candidate translations:

- Mark a phrase in the candidate as an error
- Suggest a new translation candidate
- Vote candidates up or down

Identifying likely errors and voting for candidates are things monolinguals can do reasonably well: even without knowing the intended interpretation, you can often identify when some part of a sentence doesn't make sense, or when one sentence seems more fluent or plausible than another. Sometimes rather than identifying errors, it is easier to suggest an entirely new translation candidate based on the information available on the target side, a variant of monolingual post-editing (Callison-Burch et al., 2004).

Any new translation candidates are then back-translated into Haitian Creole, and any spans marked as translation errors are projected back to identify the corresponding spans in the source sentence, using word alignments as the bridge (cf. Hwa et al. (2002), Yarowsky et al. (2001)).[2] The Haitian Creole speakers can then:

- Rephrase the entire source sentence (cf. (Morita and Ishida, 2009))
- "Explain" spans marked as errors
- Vote candidates up or down (based on the back-translation)

Source speakers can "explain" error spans by offering a different way of phrasing that piece of the source sentence (Resnik et al., 2010), in order to produce a new source sentence, or by annotating the spans with images (e.g. via Google image search) or Web links (e.g. to Wikipedia). The protocol then continues: new source sentences created via partial- or full-sentence paraphrase pass back through MT to the English side, and any explanatory annotations are projected back to the corresponding spans in the English candidate translations (where the error spans had been identified). The process is asynchronous: participants on the Haitian Creole and English sides can work independently on whatever is available to them at any time. At any point, the voting-based scores can be used to extract a 1-best translation.

In summary, the MonoTrans2 framework uses noisy MT to cross the language barrier, and supports monolingual participants in doing small tasks that gain leverage from redundant information, the human capacity for linguistic and real-world inference, and the wisdom of the crowd.

3 Experiment

We recruited 26 English speakers and 4 Haitian Creole speakers. The Haitian Creole speakers were recruited from Haiti and do not speak English. Five of the 26 English speakers were paid UMD undergraduates; the other 21 were volunteer researchers, graduate students, and staff unrelated to this research. [3] Over a 13 day period, Haitian Creole and English speaker efforts totaled 15 and 29 hours, respectively.

4 Data Sets

Our original goal of fully processing the entire SMS clean test and devtest sets could not be realized in the available time, owing to unanticipated reshuffling of the data by the shared task organizers and logistical challenges working with participants in Haiti. Table 1 summarizes the data set sizes before and after reshuffling. We put 1,224 sentences from the pre-reshuffling test set, interspersed with 123 of the 925 sentences from the pre-reshuffling devtest set, into the system — 1,347 sentences in total. We report

	before	after
test	1,224	1,274
devtest	925	900

Table 1: SMS clean data sets before and after reshuffling

[1] For the work reported here, we used Google Translate as the MT component via the Google Translate Research API.

[2] The Google Translate Research API provides alignments with its hypotheses.

[3] These, obviously, did not include any of the authors.

results on the union of pre- and post-reshuffling devtest sentences (Set A, $|A| = 1516$), and the post-reshuffling test set (Set B, $|B| = 1274$).

5 Evaluation

Of the 1,347 sentences available for processing in MonoTrans2, we define three subsets:

- $Touched$: Sentences that were processed by at least one person (657 sentences)
- *Each-side*: Sentences that were processed by at least one English speaker followed by at least one Haitian Creole speaker (431 sentences)
- $Full$: Sentences that have at least three translation candidates, of which the most voted-for one received at least three votes (207 sentences)

We intersect these three sets with sets A and B in order to evaluate MonoTrans2 output against the provided references (Table 2).[4]

Set S	$\|S\|$	$\|S \cap A\|$	$\|S \cap B\|$
$Touched$	657	162	168
Each-side	431	127	97
$Full$	207	76	60

Table 2: Data sets for evaluation and their sizes

Tables 3 and 4 report two automatic scoring metrics, uncased BLEU and TER, comparing MonoTrans2 (M2) against Google Translate (GT) as a baseline.

Set	Condition	BLEU	TER
$Touched \cap A$	GT	21.75	56.99
	M2	23.25	57.27
Each-side $\cap A$	GT	21.44	57.51
	M2	21.47	58.98
$Full \cap A$	GT	25.05	54.15
	M2	27.59	52.78

Table 3: BLEU and TER results for different levels of completion on the devtest set A

Since the number of sentences in each evaluated set is different (Table 2), we cannot directly compare scores between the sets. However, Table 4 shows that when the MonoTrans2 process is run on test items "to completion", in the sense defined by "Full" (i.e. $Full \cap B$), we see a dramatic BLEU gain of 6.39, and a drop in TER of 2.29 points. Moreover, even when only target-side or only source-side monolingual participation is available we see a gain of 4.31 BLEU and a drop of 1.73 TER points ($Touched \cap B$).

Set	Condition	BLEU	TER
$Touched \cap B$	GT	19.78	59.88
	M2	24.09	58.15
Each-side $\cap B$	GT	21.15	56.88
	M2	23.80	57.19
$Full \cap B$	GT	22.51	54.51
	M2	28.90	52.22

Table 4: BLEU and TER results for different levels of completion on the test set B

By contrast, the results on the devtest data are encouraging, but arguably mixed (Table 3). In order to step away from the vagaries of single-reference automatic evaluations, therefore, we also conducted an evaluation based on human judgments. Two native English speakers unfamiliar with the project were recruited and paid for fluency and adequacy judgments: for each target translation paired with its corresponding reference, each evaluator rated the target sentence's fluency and adequacy on a 5-point scale, where fluency of 5 indicates complete fluency and adequacy of 5 indicates complete preservation of meaning (Dabbadie et al., 2002).[5]

Sentences	N	Google		MonoTrans2	
$Full \cap A$	76	18	(24%)	30	(39%)
$Full \cap B$	60	15	(25%)	23	(38%)

Table 5: Number of sentences with maximum possible adequacy (5) in $Full \cap A$ and $Full \cap B$, respectively.

Similar to Hu et al. (2011), we adopt the very conservative criterion that a translation output is considered correct only if *both* evaluators independently give it a rating of 5. Unlike Hu et al. (2011), for whom children's book translation requires both fluency and adequacy, we make this a requirement only

[4] Note that according to these definitions, $Touched$ contains both *Each-side* and $Full$, but *Each-side* does not contain $Full$.

[5] Presentation order was randomized.

for adequacy, since in this scenario what matters to aid organizations is not whether a translation is fully fluent, but whether it is correct. On this criterion, the Google Translate baseline of around 25% correct improves to around 40% for Monotrans, consistently for both the devtest and test data (Table 5). Nonetheless, Figures 1 and 2 make it clear that the improvements in fluency are if anything more striking.

5.1 Statistical Analysis

Variable	Adequacy	Fluency
Positive		
mostSingleCandidateVote	**	***
candidateCount	**	**
numOfAnswers	*	NS
Negative		
roundTrips	***	***
voteCount	*	.

Table 6: Effects of independent variables in linear regression for 330 touched sentences
(Signif. codes: '***' 0.001, '**' 0.01, '*' 0.05, '.' 0.1)

In addition to the main evaluation, we investigated the relationship between tasks performed in the MonoTrans2 system and human judgments using linear regression and an analysis of variance. We evaluate the set of all 330 touched sentences in $Touched \cap A$ and $Touched \cap B$ in order to understand which properties of the MonoTrans2 process correlate with better translation outcomes.

Our analysis focused on improvement over the Google Translate baseline, looking specifically at the improvement based on the human evaluators' averaged fluency and adequacy scores.

Table 6 summarizes the positive and negative effects for five of six variables we considered that came out significant for at least one of the measures. [6]

The positive results were as expected. Having more votes for the winning candidate (*mostSingleCandidateVote*) made it more successful, since this means that more people felt it was a good representative translation. Having more candidates to choose from (*candidateCount*) meant that more people had taken the time to generate alternatives, reflecting attention paid to the sentence. Also, the amount of attention paid to target speakers' requests for clarification (*numOfAnswers*) is as expected related to the adequacy of the final translation, and perhaps as expected does not correlate with fluency of the output since it helps with meaning and not actual target-side wording.

We were, however, confused at first by the negative influence of the *roundTrips* measure and *voteCount* measures. We conjecture that the first effect arises due to a correlation between roundTrips and translation difficulty; much harder sentences would have led to many more paraphrase requests, and hence to more round trips. We attempted to investigate this hypothesis by testing correlation with a naive measure of sentence difficulty, length, but this was not fruitful. We suspect that inspecting use of abbreviations, proper nouns, source-side mistakes, and syntactic complexity would give us more insight into this issue.

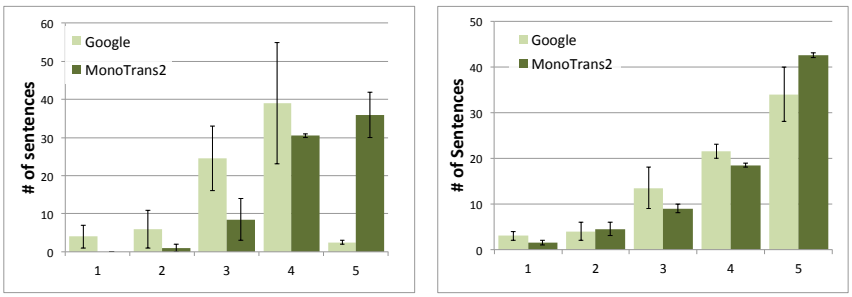
As for *voteCount*, the negative correlation is understandable when considered side by side with the other vote-based measure, *mostSingleCandidateVote*. Having a higher number of votes for the winning candidate leads to improvement (strongly significant for both adequacy and fluency), so a higher general vote count means that people were also voting more times for other candidates. Hence, once the positive winning vote count is taken into account, the remaining votes actually represent disagreement on the candidates, hence correlating negatively with overall improvement over baseline.

It is important to note that when these measures are all considered together, they show that there is a clear correlation between the MonoTrans2 system's human processing and the eventual increase in both quality and fluency of the sentences. As people give more attention to sentences, these sentences show better performance, as judged by increase over baseline.

6 Discussion

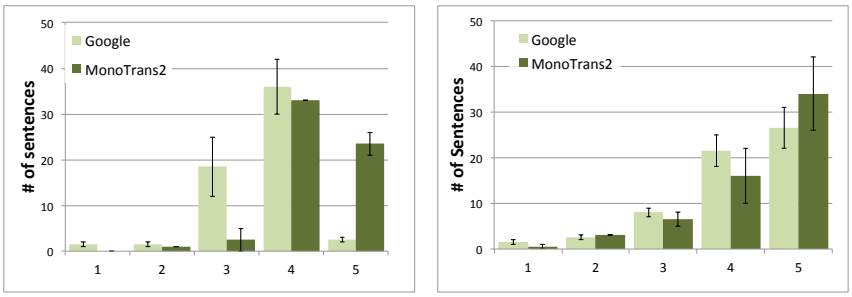
Our experiment did not address acquisition of, and incentives for, monolingual participants. In fact, getting time from Haitian Creole speakers, even for pay,

[6] A sixth, *numOfVoters*, was not significant in the linear regression for either adequacy or fluency.



(a) Fluency Distribution (b) Adequacy Distribution

Figure 1: Human judgments for fluency and adequacy in fully processed devtest items ($Full \cap A$)



(a) Fluency Distribution (b) Adequacy Distribution

Figure 2: Human judgments for fluency and adequacy in fully processed test items ($Full \cap B$)

created a large number of logistical challenges, and was a contributing factor as to why we did not obtain translations for the entire test set. However, availability of monolingual participants is not the issue being addressed in this experiment: we are confident that in a real-world scenario like the Haitian or Japanese earthquakes, large numbers of monolingual volunteers would be eager to help, certainly in larger total numbers than *bilingual* volunteers. What matters here, therefore, is not how much of the test set was translated in total, but how much the translations improved for the sentences where monolingual crowdsourcing was involved, compared to the MT baseline, and what throughput might be like in a real-world scenario.

We also were interested in throughput, particularly in comparison to bilingual translators. In previous experimentation (Hu et al., 2011), throughput in MonoTrans2 extrapolated to roughly 800 words per day, a factor of 2.5 slower than professional translators' typical speed of 2000 words per day. In this experiment, overall translation speed averaged about 300 words per day, a factor of more than 6 times slower. However, this is an extremely pessimistic estimate, for several reasons. First, our previous experiment had more than 20 users per side, while here our Haitian crowd consisted of only four people. Second, we discovered after beginning the experiment that the translation of our instructions into Haitian Creole had been done somewhat sloppily. And, third, we encountered a range of technical and logistical problems with our Haitian participants, ranging from finding a location with Internet access to do the work (ultimately an Internet Café turned out to be the best option), to slow and sporadic connections (even in an Internet Café), to relative lack of motivation for part-time rather than full-time work. It is fair to assume that in a real-world scenario, some unanticipated problems like these might crop up, but it also seems fair to assume that many would not; for example, most people from the Haitian Creole and French-speaking communities who volunteered using Munro et al.'s system in January 2010 were not themselves located in the

third world.

Finally, regarding quality, the results here are promising, albeit not as striking as those Hu et al. (2011) obtained for Spanish-German translation of children's books. The nature of SMS messages themselves may have been a contributing factor to the lower translation adequacy: even in clean form, these are sometimes written using shorthand (e.g. "SVP"), and are sometimes not syntactically correct. The text messages are seldom related to each other, unlike sentences in larger bodies of text where even partially translated sentences can be related to each other to provide context, as is the case for children's books. One should also keep in mind that the underlying machine translation engine, Google Translate between Haitian Creole and English, is still in an alpha phase.

Those considerations notwithstanding, it is encouraging to see a set of machine translations get better without the use of any human bilingual expertise. We are optimistic that with further refinements and research, monolingual translation crowdsourcing will make it possible to harness the vast number of technologically connected people who want to help in some way when disaster strikes.

7 Acknowledgments

This research is supported by NSF contract #BCS0941455 and by a Google Research Award.

References

Olivia Buzek, Philip Resnik, and Benjamin B. Bederson. 2010. Error driven paraphrase annotation using mechanical turk. In *NAACL 2010 Workshop on Creating Speech and Text Language Data With Amazon's Mechanical Turk*.

Chris Callison-Burch, Colin Bannard, , and Josh Schroeder. 2004. Improving statistical translation through editing. In *Workshop of the European Association for Machine Translation*.

Marianne Dabbadie, Anthony Hartley, Margaret King, Keith J. Miller, Widad Mustafa El Hadi, Andrei Popescu-Belis, Florence Reeder, and Michelle Vanni. 2002. A hands-on study of the reliability and coherence of evaluation metrics. In *Workshop at the LREC 2002 Conference*, page 8. Citeseer.

Chang Hu, Benjamin B. Bederson, and Philip Resnik. 2010. Translation by iterative collaboration between monolingual users. In *Proceedings of Graphics Interface 2010 on Proceedings of Graphics Interface 2010*, pages 39–46, Ottawa, Ontario, Canada. Canadian Information Processing Society.

Chang Hu, Ben Bederson, Philip Resnik, and Yakov Kronrod. 2011. Monotrans2: A new human computation system to support monolingual translation. In *Human Factors in Computing Systems (CHI 2011)*, Vancouver, Canada, May. ACM, ACM.

Chang Hu. 2009. Collaborative translation by monolingual users. In *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*, pages 3105–3108, Boston, MA, USA. ACM.

Rebecca Hwa, Philip Resnik, Amy Weinberg, and Okan Kolak. 2002. Evaluating translational correspondence using annotation projection. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 392–399, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Daisuke Morita and Toru Ishida. 2009. Designing protocols for collaborative translation. In *PRIMA '09: Proceedings of the 12th International Conference on Principles of Practice in Multi-Agent Systems*, pages 17–32, Berlin, Heidelberg. Springer-Verlag.

Robert Munro. 2010. Crowdsourced translation for emergency response in haiti: the global collaboration of local knowledge. In *AMTA Workshop on Collaborative Crowdsourcing for Translation*. Keynote.

Alexander J. Quinn, Bederson, and Benjamin B. Bederson. 2011. Human computation: A survey and taxonomy of a growing field. In *Human Factors in Computing Systems (CHI 2011)*, Vancouver, Canada, May. ACM, ACM.

Philip Resnik, Olivia Buzek, Chang Hu, Yakov Kronrod, Alexander J. Quinn, and Benjamin B. Bederson. 2010. Improving translation via targeted paraphrasing. In *EMNLP*.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *HLT '01: Proceedings of the first international conference on Human language technology research*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.