Using Targeted Paraphrasing and Monolingual Crowdsourcing to Improve Translation

PHILIP RESNIK, OLIVIA BUZEK, YAKOV KRONROD, CHANG HU,
ALEXANDER J. QUINN, and BENJAMIN B. BEDERSON, University of Maryland

Targeted paraphrasing is a new approach to the problem of obtaining cost-effective, reasonable quality translation, which makes use of simple and inexpensive human computations by monolingual speakers in combination with machine translation. The key insight behind the process is that it is possible to spot likely translation errors with only monolingual knowledge of the target language, and it is possible to generate alternative ways to say the same thing (i.e., paraphrases) with only monolingual knowledge of the source language. Formal evaluation demonstrates that this approach can yield substantial improvements in translation quality, and the idea has been integrated into a broader framework for monolingual collaborative translation that produces fully accurate, fully fluent translations for a majority of sentences in a real-world translation task, with no involvement of human bilingual speakers.

Categories and Subject Descriptors: I.2.7 [**Natural Language Processing**]: Machine translation

General Terms: Design, Experimentation, Human Factors, Languages

Additional Key Words and Phrases: Monolingual, paraphrase, machine translation, translation, translation interface, human computation, wisdom of crowds, crowdsourcing

ACM Reference Format:
Resnik, P., Buzek, O., Kronrod, Y., Hu, C., Quinn, A. J., and Bederson, B. B. 2013. Using targeted paraphrasing and monolingual crowdsourcing to improve translation. ACM Trans. Intell. Syst. Technol. 4, 3, Article 38 (June 2013), 21 pages.
DOI: http://dx.doi.org/10.1145/2483669.2483671

1. INTRODUCTION

For most of the world's languages, the availability of translation is limited to two possibilities: high quality at high cost, via professional translators, and low quality at low cost, via Machine Translation (MT). The spectrum between these two extremes is very poorly populated, and at any point on the spectrum the ready availability of translation is limited to only a small fraction of the world's languages. There is, of course, a long history of technological assistance to translators, improving cost effectiveness using translation memory [Laurian 1984; Bowker and Barlow 2004] or other interactive tools to assist translators [Esteban et al. 2004; Khadivi et al. 2006]. And there is a recent and rapidly growing interest in crowdsourcing with nonprofessional translators, which can sometimes be remarkably effective [Huberdeau et al. 2008;

This work has been supported in part by the National Science Foundation under awards BCS0941455 and IIS0838801 and by a Google Research Award.
Authors' addresses: P. Resnik (corresponding author), Department of Linguistics, University of Maryland, College Park, MD; email: resnik@umd.edu; O. Buzek, Department of Computer Science, University of Maryland, College Park, MD; Y. Kronrod, Department of Linguistics, University of Maryland, College Park, MD; C. Hu, A. J. Quinn, B. B. Bederson, Department of Computer Science, University of Maryland, College Park, MD.

DOI: http://dx.doi.org/10.1145/2483669.2483671

Munro 2010; Hester et al. 2010; Meedan 2011; Twitter 2011; TED 2011; Facebook 2011; 99Translations 2011; GetLocalization 2011; Zaidan and Callison-Burch 2011]. However, all these alternatives face a central availability bottleneck: they require the participation of humans with bilingual expertise.

In this article, we report on a new exploration of the middle ground, taking advantage of a virtually unutilized resource: speakers of the source and target language who are *effectively monolingual*, that is, who each only know one of the two languages relevant for the translation task. The solution we are proposing has the potential to provide a more cost-effective approach to translation in scenarios where machine translation *would* be considered acceptable to use, if only it were generally of high enough quality. This would clearly exclude tasks like translation of medical reports, business contracts, or literary works, where the validation of a qualified bilingual translator is absolutely necessary. However, it does include a great many real-world scenarios, such as following news reports in another country, reading international comments about a product, or generating a decent first draft translation of a Wikipedia page for Wikipedia editors to improve. It also has the potential to vastly reduce the burden of human effort for use cases in which bilingual translators post-edit machine translation output [Guerra 2003].

We call the technique described here *targeted paraphrasing*.[1] In a nutshell, target-language monolinguals identify parts of an initial machine translation that don't appear to be right, and source-language monolinguals provide the MT system with alternative phrasings that might lead to better translations; these are then passed through MT again and the best scoring hypothesis is selected as the final translation.

The use of monolingual participants in a human-machine translation process is not entirely new. Callison-Burch et al. [2004] pioneered the exploration of monolingual post-editing within the MT community, an approach extended more recently to provide richer information to the user by Albrecht et al. [2009] and Koehn [2009]. Shahaf and Horvitz [2010] use machine translation as a specific instance of a general game-based framework for combining a range of machine and human capabilities. There have also been at least two independently developed human-machine translation frameworks that employ an iterative protocol involving monolinguals on both the source and target side. In the first of these, Morita and Ishida [2009] describe a system in which target and source language speakers perform editing of MT output to improve fluency and adequacy, respectively; they utilize source-side paraphrasing at a course-grain level, although their approach is limited to requests to paraphrase the entire sentence when the translation cannot be understood. The second of these provides the broader context for the work reported here: we have independently developed a protocol similar in spirit to that of Morita and Ishida, in which cross-language communication is enhanced by metalinguistic communication in the user interface [Bederson et al. 2010; Hu et al. 2011]. The technique we describe in this article can be viewed as compatible with the richer protocol- and game-based approaches, but it is considerably simpler.

In Sections 2 through 5 we describe our method and present evaluation results on Chinese-English translation. Unlike other work on translation using monolingual human participants, the technique we present here also offers clear opportunities to replace human participation with machine components if the latter are up to the task;

[1]This article revises and significantly extends work first introduced in earlier publications [Buzek et al. 2010; Resnik et al. 2010], providing clearer and more detailed presentations of prior results, extending preliminary Chinese-English experimentation to the full test set (Section 4), adding more extensive analysis and additional experimentation (see in particular Section 5), and briefly summarizing the results of Hu et al. [2011] in order to provide a broader context for the work, where targeted paraphrasing is part of a collaborative interface that permits a wider variety of monolingual tasks (Section 7). A description of our work on monolingual translation crowdsourcing, along with an up-to-date list of publications, is maintained at http://www.cs.umd.edu/hcil/monotrans/.

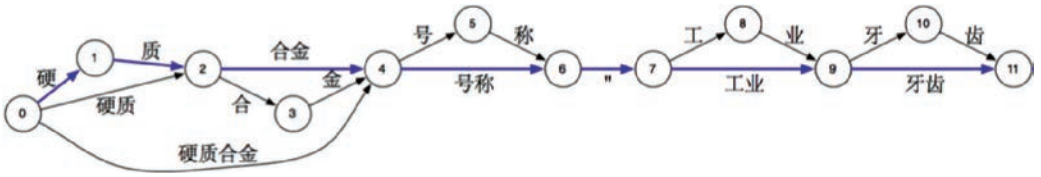

Fig. 1. Illustration of subsentential alternatives expressing the same meaning: a lattice containing multiple segmentations of the same Chinese input.

we present promising results along these lines in Section 6. In Section 7 we briefly summarize the collaborative translation protocol that provides the broader context for this work, including evaluation results [Hu et al. 2011], before wrapping up in Section 8 with conclusions and directions for future work.

2. TARGETED PARAPHRASING

The starting point for the work we report on in this article is an observation: the source sentence provided as input to an MT system is just one of many ways in which the meaning could have been expressed, and for any given MT system, some forms of expression are easier to translate than others. The same basic observation has been applied quite fruitfully over the past several years to deal with statistical MT challenges involving segmentation, morphological analysis, and more recently, source language word order [Dyer 2007; Dyer et al. 2008, 2010; Dyer and Resnik 2010]. For example, Figure 1 shows a source translation lattice that contains alternative segmentations of a Chinese input sentence. A decoder enabled to handle lattice input, for example, Dyer et al. [2010], can exploit inputs of this kind by identifying the path through the lattice that leads to the best scoring translation hypothesis, making *subsentential* choices as to which segmentation best contributes to a good hypothesis score. For example, given the input in the figure, a translation model lacking a good translation for the first Chinese word in its entirety (spanning nodes 0 to 4) could instead choose to traverse arcs 0–1, 1–2, and 2–4, taking advantage of better translation possibilities for the smaller translation units.

Here we apply the same core idea—providing a wider range of subsentential alternatives for source language phrases—not at the level of segmentation of morphological analysis, but to the surface expression of meaning. For example, consider the following real example of translation from English to French by an automatic MT system.

—*Source.* Polls indicate Brown, a state senator, and Coakley, Massachusetts' Attorney General, are locked in a virtual tie to fill the late Sen. Ted Kennedy's Senate seat.
—*System.* Les sondages indiquent Brown, un sénateur d'état, et Coakley, Massachusetts' Procureur général, sont enfermés dans une cravate virtuel à remplir le regretté sénateur Ted Kennedy's siège au Sénat.

A French speaker can look at this automatic translation and see immediately that the underlined parts are wrong, even without knowing the intended source meaning. We can identify the spans in the source English sentence that are responsible for these badly translated French spans, and change them to alternative expressions with the same meaning (e.g., changing *Massachusetts' Attorney General* to *the Attorney General of Massachusetts*); if we do so and then use the same MT system again, we obtain a translation that is still imperfect (e.g., *cravate* means necktie), but is more acceptable.

—*System.* Les sondages indiquent que Brown, un sénateur d'état, et Coakley, le procureur général du Massachusetts, sont enfermés dans une cravate virtuel à pourvoir le siége au Sénat de Sen. Ted Kennedy, qui est décédé récemment.

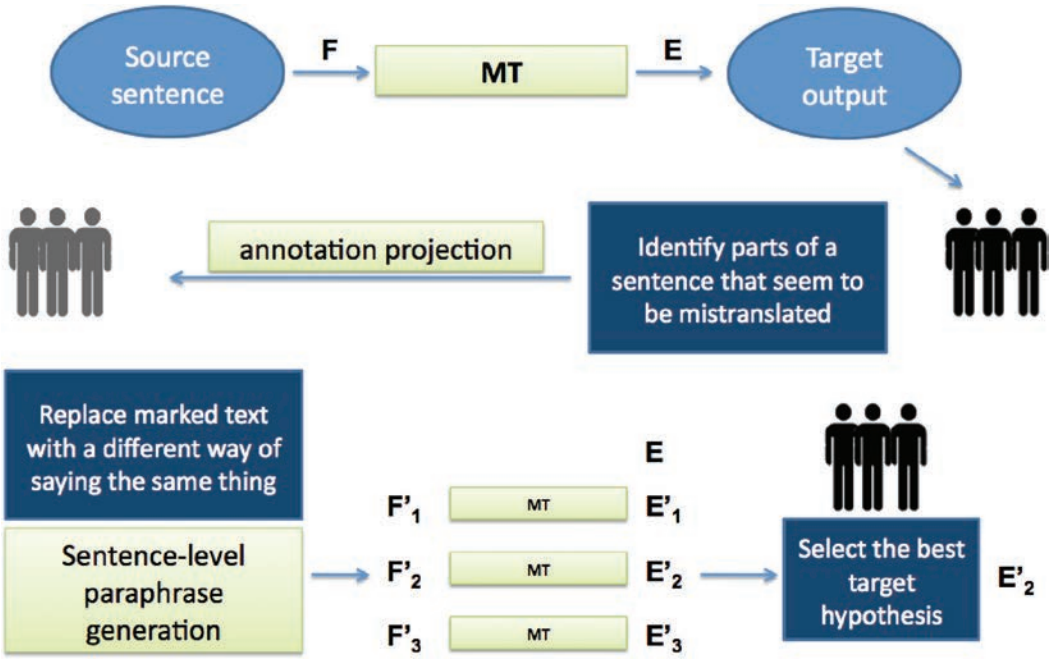

Fig. 2. Visual summary of the targeted paraphrase process.

Operationally, then, translation with targeted paraphrasing includes the following steps (Figure 2).

Initial machine translation. For this article, we use the Google Translate Research API, which, among other advantages, provides word-level alignments between the source text and its output. In principle, however, any automatic translation system can be used in this role, potentially at some cost to quality, by performing post hoc target-to-source alignment.

Identification of mistranslated spans. This step identifies parts of the source sentence that lead to ungrammatical, nonsensical, or apparently incorrect translations on the target side. In the experiments of Sections 3 and 4, this step is performed by having monolingual target speakers identify likely error spans on the target side, as in the French example before, and projecting those spans back to the source spans that generated them using word alignments as the bridge [Hwa et al. 2005; Yarowsky et al. 2001]. In Section 6, we describe a heuristic but effective method for performing this fully automatically. Du et al. [2010] explore the use of source paraphrases without targeting apparent mistranslations, using lattice translation [Dyer et al. 2008] to efficiently represent and decode the resulting very large space of paraphrase alternatives.

Source paraphrase generation. This step generates alternative expressions for the source spans identified in the previous step. In this article, it is performed by monolingual source speakers who perform the paraphrase task: the speaker is given a sentence with a phrase span marked, and is asked to replace the marked text with a different way of saying the same thing, so that the resulting sentence still makes sense and means the same thing as the original sentence. To illustrate in English, someone seeing *John and Mary took a European vacation this summer* might supply the paraphrase *Mary went on a European*, verifying that the resulting *John and Mary*

went on a European vacation this summer preserves the original meaning. This step can also be fully automated [Max 2009] by taking advantage of bilingual phrase-table pivoting [Bannard and Callison-Burch 2005]; see Max [2010] for a related approach in which the paraphrases of a source phrase are used to refine the estimated probability distribution over its possible target phrases.

Generating sentential source paraphrases. For each sentence, there may be multiple paraphrased spans. These are multiplied out to provide full-sentence paraphrases. For example, if two nonoverlapping source spans are each paraphrased in three ways, we generate 9 sentential source paraphrases, each of which represents an alternative way of expressing the original sentence.[2]

Machine translation of alternative sentences. The alternative source sentences, produced via paraphrase, are sent through the same MT system.

Hypothesis selection. A single-best translation hypothesis is selected, for example, on the basis of the translation system's model score. In principle, one could also combine the alternatives into a lattice representation and decode to find the best path using lattice translation [Dyer et al. 2008]; see Du et al. [2010]. One can also present translation alternatives to a target speaker for selection, similarly to Callison-Burch et al. [2004].

Notice that with the exception of the initial translation, each remaining step in this pipeline can involve either human participation or fully automatic processing. The targeted paraphrasing framework therefore defines a rich set of intermediate points on the spectrum between fully automatic and fully human translation, of which we explore only a few in this article.

3. PILOT STUDY

In order to assess the potential of our approach, we conducted a small pilot study, using eleven sentences in simplified Chinese selected from the article on "Water" in Chinese Wikipedia (http://zh.wikipedia.org/zh-cn/%E6%B0%B4). This article was chosen because its topic is well known in both English-speaking and Chinese-speaking populations. The first five sentences were taken from the first paragraph of the article. The other six sentences were taken from a randomly chosen paragraph in the article. As a preprocessing step, we removed any parenthetical items from the input sentences, such as "(H_20)". The shortest sentence in this set has 12 Chinese characters, the longest has 54.[3]

Human participation in this task was accomplished using Amazon Mechanical Turk, an online marketplace that enables human performance of small Human Intelligence Tasks (HITs) in return for micropayments. For each sentence, after we translated it automatically (using Google Translate), three English-speaking Mechanical Turk workers ("Turkers") on the target side performed identification of mistranslated spans. Each span identified was projected back to its corresponding source span, and three Chinese-speaking Turkers were asked to provide paraphrases of each source span. These tasks were easy to perform (no more than around 30 seconds to complete on average) and inexpensive (less than $1 for the entire pilot study).[4] The Chinese source span paraphrases were then used to construct full-sentence paraphrases, which were

[2]If spans overlap, one of them is chosen at random and the others discarded, to create a nonoverlapping set.
[3]Note that this page is not a translation of the corresponding English Wikipedia page or vice versa.
[4]The four English-speaking Turkers were recruited through the normal Mechanical Turk mechanism. The three Chinese-speaking Turkers were recruited offline by the authors in order to quickly obtain results, although they participated as full-fledged Turkers.

retranslated, once again by Google Translate, to produce the output of the targeted paraphrasing translation process.

The initial translation outputs from Google Translate (GT) and the results of the targeted paraphrasing translation process (TP) were evaluated according to widely used criteria of fluency and adequacy.[5] Fluency ratings were obtained on a 5-point scale from three native English speakers without knowledge of Chinese. Translation adequacy ratings were obtained from three native Chinese speakers who are also fluent in English; they assessed adequacy of English sentences by comparing the communicated meaning to the Chinese source sentences.

Fluency was rated on the following scale.

(1) Unintelligible: nothing or almost nothing of the sentence is comprehensible.
(2) Barely intelligible: only a part of the sentence (less than 50%) is understandable.
(3) Fairly intelligible: the major part of the sentence passes.
(4) Intelligible: all the content of the sentence is comprehensible, but there are errors of style and/or of spelling, or certain words are missing.
(5) Very intelligible: all the content of the sentence is comprehensible. There are no mistakes.

Adequacy was rated on the following scale.

(1) None of the meaning expressed in the reference sentence is expressed in the sentence.
(2) Little of the reference sentence meaning is expressed in the sentence.
(3) Much of the reference sentence meaning is expressed in the sentence.
(4) Most of the reference sentence meaning is expressed in the sentence.
(5) All meaning expressed in the reference sentence appears in the sentence.

For each GT output, we averaged across the ratings of the alternative TP to produce average TP fluency and adequacy scores. The average GT output ratings, measuring the pure machine translation baseline, were 2.36 for fluency and 2.91 for adequacy. Averaging across the TP outputs, these rose to 3.32 and 3.49, respectively.

One could argue that a more sensible evaluation is not to average across alternative TP outputs, but rather to simulate the behavior of a target-language speaker who simply chooses the one translation among the alternatives that seems most fluent. If we select the most fluent TP output for each source sentence according to the English speakers' average fluency ratings, we obtain average test set ratings of 3.58 for fluency and 3.73 for adequacy. These are respective gains of 0.82 and 1.21 over the baseline initial MT output, each on a 5-point scale (Figure 3).

Figure 4 shows a selection of outputs: we present the two cases where the most fluent TP alternative shows the greatest gain in average fluency rating (best gain +2.67); two cases near the median gain in average fluency (median +1); and the worst two cases with respect to effect on average fluency rating (worst −0.33). The table accurately conveys a qualitative impression corresponding to the quantitative results: the overall quality of translations appears to be improved by our process consistently, despite the absence of any bilingual input in the improvements.

4. CHINESE-ENGLISH EVALUATION

Encouraged by our pilot study, we conducted a more extensive evaluation using Chinese-English test data taken from the NIST MT'08 machine translation evaluation, in order to obtain fully automatic translation evaluation scores. The NIST MT'08

[5]Fluency and adequacy are long standing, widely accepted human measures of translation quality [Doyon et al. 1998; Dabbadie et al. 2002]. An anonymous reviewer correctly points out that, more recently, human assessments of preference (which of these translations do you like better?) have joined adequacy as a benchmark for evaluation [Przybocki et al. 2009; Callison-Burch et al. 2010]. We plan to include preference judgments in future work.

Fig. 3. Summary of pilot study results: strong gains for average ratings of fluency and adequacy (preservation of meaning in translation).

Condition	Fluency	Adequacy	Sentence
GT	1.33	2.33	Water play life evolve into important to use.
TP	4.00	4.33	Water in the evolution of life played an important role.
GT	1.33	2.67	Human civilization from the source of the majority of large rivers in the domain.
TP	3.33	4.67	Most of the origin of human civilization in river basin.
GT	2.33	3.00	In human daily life, the water in drinking, cleaning, washing and other side to make use of an indispensable.
TP	3.67	3.33	In human daily life, water for drinking, cleaning, washing and other essential role.
GT	2.00	2.33	Eastern and Western ancient Pak prime material view of both the water regarded as a kind of basic groups into the elements, water is the Chinese ancient five rows of a; the West ancient four elements that also have water.
TP	3.00	3.33	East and West in ancient concept of simple substances regarded water as a basic component elements. Among them, the five elements of water is one of ancient China; Western ancient four elements that also have water.
GT	4.00	4.00	Early cities will generally be in the water side of the establishment, in order to solve irrigation, drinking and sewage problems.
TP	4.67	4.33	Early cities are generally built near the water to solve the irrigation, drinking and sewage problems.
GT	3.0	3.33	Human very early on began to produce a water awareness.
TP	2.67	3.00	Man long ago began to understand the water produced.

Fig. 4. Original Google Translate output (GT) for the pilot study in Section 3, together with translations produced by the targeted paraphrase translation process (TP), selected to show a range from strong to weak improvements in fluency.

test set contains 1,357 test sentences. These underwent the same targeted paraphrasing process as in the pilot study, with the addition of a basic step to filter out cheaters: we disregarded as invalid any responses consisting purely of ASCII characters (signifying a non-Chinese response) or responses that were identical to the original source text. (The issue of cheating is discussed further in Section 5.) Target English speakers identified 4292 potential mistranslation spans, or 3.17 spans per sentence, that yielded at least one source paraphrase on the source Chinese side. Chinese speakers provided 1513 valid paraphrases. This process produced 649 sentences with both valid identification of error spans and valid proposals for source-side paraphrase alternatives.

Condition	BLEU
Google Translate (GT, baseline)	28.33
Google Translate (GT) n-best oracle	28.47
Targeted paraphrasing (TP) one-best	30.01
Targeted paraphrasing (TP) oracle	30.79
Human upper bound	49.41

Fig. 5. Results on a 49-sentence subset of the NIST MT'08 Chinese-English test set.

The entire cost for the human tasks in this experiment was $408.386, or a bit under $0.30 per sentence in the test set on average.[6]

In order to get a sense of improvements and how much improvement is possible, using the proposed technique, we computed preliminary results on a subset of 49 sentences obtained before the full crowdsourcing process was complete. Figure 5 summarizes an evaluation in standard fashion using the BLEU evaluation metric [Papineni et al. 2002], evaluating against the four English MT'08 references for each Chinese sentence and using one-best output from Google Translate (GT) as the baseline.

Since the targeted paraphrasing translation process (TP) produces multiple hypotheses—one automatic translation output per sentential paraphrase—we selected the one-best output for each sentence by selecting the highest scoring English translation, according to the translation score delivered with each output by the Google Translate Research API. (The original translation was, of course, included among the candidates for selection.) This achieves an improvement of 1.68 BLEU points over baseline on the 49-sentence test set (listed as Targeted paraphrasing (TP) one-best in the figure).

One could argue that this outcome is simply a result of having more hypotheses to choose from, not a result of the targeted paraphrasing process itself. In order to rule out this possibility, we generated $(n + 1)$-best Google translations, setting n for each sentence to match the number of alternative translations generated via targeted paraphrasing. We then chose the best translation for each sentence, among the $(n+1)$-best Google hypotheses, via oracle selection, using the TERp metric [Snover et al. 2009] to score each hypothesis in the $(n+1)$-best list against the reference translations.[7] The resulting BLEU score for the full set showed negligible improvement (Google Translate (GT) n-best oracle).[8]

We did a similar oracle-best calculation for targeted paraphrasing (Targeted Paraphrasing (TP) oracle). Here, instead of picking the best targeted paraphrasing output via the Google Translate score, as we did for Targeted Paraphrasing (TP) one-best, we evaluated those outputs against the reference translations using TERp [Snover et al. 2009], picking the one with the best score as the oracle translation, exactly as was done for Google Translate (GT) n-best oracle. The result shows a potential gain of 2.46 BLEU points over the baseline, selecting the oracle translation in this way.

[6]Invalid paraphrase responses were rejected, that is, zero cost.

[7]Translation Error Rate (TER) [Snover et al. 2006] and Translation-Edit-Rate Plus (TERp) [Snover et al. 2009], like BLEU, are translation evaluation metrics that compare candidate translations against collections of (presumed-correct) human reference translations. Unlike BLEU, which is based on *n*-gram precision, TER and TERp are generalizations of string edit distance. TERp goes beyond TER in a number of ways, among them permitting string-internal substitutions that can include synonyms and paraphrases. Each metric has its own advantages and disadvantages. We selected among them as appropriate for the specific task.

[8]An "oracle" telling us which variant is best is not available in the real world, of course, but in situations like this one, oracle studies are often used to establish the magnitude of the potential gain [Och et al. 2004]. An anonymous reviewer suggests considering a "lattice oracle", that is, scoring the path in a translation lattice that has the lowest error rate relative to the references; we plan to do so once we move our experimental baseline from Google Translate, which will not conveniently produce lattice output, to the *cdec* decoder [Dyer et al. 2010], which will.

Error Spans	# Sentences	Google (BLEU)	ParaTrans (BLEU)	Delta
1	183	27.94	28.72	0.78
2	206	29.39	31.10	1.71
3	161	28.18	29.16	0.98
4	123	24.63	28.40	3.77
5	101	29.59	31.22	1.63
6	51	25.38	26.44	1.06
7	58	24.42	26.68	2.26

Fig. 6. Results improving Google Translation output on NIST MT'08 data using targeted paraphrase.

In addition to aggregate evaluation using BLEU, we also looked at oracle results on a per-sentence basis using TERp (since BLEU is known to be more appropriate to use at document level, not sentence level). Identifying the best sentential paraphrase alternative using TERp as an oracle, we find that the TERp score would improve for 32 of the 49 test sentences, 65.3%. For those 32 sentences, the average gain is 8.36 TERp points.[9] A fairer measure is the average obtained when scoring zero gain for the 17 sentences where no improvement was obtained; taking these into account, that is, assuming an oracle who chooses the original translation if none of the paraphrase-based alternatives is better, the average improvement over the entire set of 49 sentences is 5.46 TERp points.

Finally, the last line in Figure 5 shows a human upper bound computed using the reference translations via cross-validation; that is, for each of the four reference translations, we evaluate it as a hypothesized translation using the other three references as ground truth; these four scores are then averaged. The value of this upper bound is quite consistent with a bound computed similarly by Callison-Burch [2009].

This preliminary evaluation on a small subset of the full NIST MT'08 test set confirmed the qualitative impressions in Figure 4 and the subjective ratings results obtained in our pilot study in Section 3. We then moved on to a similar analysis for the full 649-sentence set.

Figure 6 presents the key results. Targeted paraphrase yielded an average improvement of +1.6 BLEU. The table breaks out these improvements by the number of error spans identified per sentence on the target side as likely to contain errors. We conjectured that there might be a "sweet spot" for the number of segments of a sentence to be paraphrased. This appears, anecdotally, to be the case, given that the 79 sentences with 4 error spans each seem to have outperformed the other sets by a substantial gain in BLEU score. In terms of the quantifiable improvements obtained for the various sentence sets, it is worth noting that gains of roughly +0.6 BLEU or higher tend to be considered meaningful by MT researchers, and gains of +1.5 BLEU are generally considered substantial.

Our oracle results establish that by taking advantage of monolingual human speakers, it is possible to obtain quite substantial gains in translation quality. Figure 7 provides a qualitative sense of how the targeted paraphrase results differ from the automatic MT output. The TP one-best results demonstrate that the majority of that oracle gain is obtained in automatic hypothesis selection, simply by selecting the paraphrase-based alternative translation with the highest translation score.

5. ADDITIONAL ANALYSIS AND THE VALUE OF QUALITY CONTROL

In the previous analysis, we saw that our approach leads to promising improvements to the baseline Google translations, even in the absence of human bilingual expertise.

[9]"Gains" refer to a lower score: since TERp is an error- or distance-based measure, lower is better.

GT: WTO chief negotiator on behalf of the United States to propose substantial reduction of agricultural subsidies, Kai Fa countries substantially reduce industrial products import tariffs to Dapo ?? Doha Round of negotiations deadlock. **TP:** World Trade Organization negotiator suggested the United States today, a substantial reduction of agricultural subsidies, developing countries substantially reduce industrial products?? Import tariffs, in order to break the deadlock in the Doha Round of trade negotiations. **REF:** the main delegates at the world trade organization talks today suggested that the us make major cuts in its agricultural subsidies and that developing countries significantly reduce import duties on industrial products in order to break the deadlock in the doha round of trade talks .
GT: Emergency session of the Palestinian prime minister Salam Fayyad state will set a new Government **TP:** Emergency session of the Palestinian Prime Minister Salam Fayyad will set the new government **REF:** state of emergency period ends ; palestinian prime minister fayyad to form new government
GT: Indian territory from south to north, one week before the start after another wet season, the provincial residents hold long drought every rain in the mood to meet the heavy rain, but did not expect rain came unexpectedly fierce, a rain disaster, roads become rivers, low-lying areas housing to make Mo in the water, transport almost paralyzed, Zhi Jin statistics about You nearly 500 people due to floods were killed. **TP:** Indian territory from south to north, one week before the start have entered into the rainy season, provincial residents hold long drought to hope rain in the mood to meet the heavy rain, but did not feed rain came unexpectedly fierce, a rain disaster, roads change the river, low-lying areas housing do not water, traffic almost to a standstill, since statistics are nearly 500 people due to floods killed. **REF:** the whole of india , from south to north , started to progressively enter the monsoon season a week ago . the residents of each state all greeted the heavy rains as relief at the end of a long drought , but didn't expect that the rain would come with unexpected violence , a real deluge . highways have become rivers ; houses in low-lying areas have been surbmerged in the water ; the transport system is nearly paralyzed . to date , figures show that nearly 500 people have unfortunately lost their lives to the floods .
GT: But the Taliban said in the meantime, the other a German hostages kidnapped in very poor health, began to fall into a coma and lost consciousness. **TP:** But the Taliban said in the meantime, another German hostages kidnapped a very weak body fell into a coma and began to lose consciousness. **REF:** but at the same time the taliban said that another german hostage who had been kidnapped was in extremely poor health , and had started to become comatose and to lose consciousness .
GT: Taliban spokesman Ahmadi told AFP in an unknown location telephone interview, said: We, through tribal elders, representatives of direct contact with South Korea. **TP:** Taliban spokesman Ahmadi told AFP in an unknown location telephone interview, said: We are through tribal elders, directly with the South Korean leadership, business **REF:** taliban spokesperson ahmadi said in a telephone interview by afp at an undisclosed location : we have established direct contact with the south korean delegation through tribal elders .

Fig. 7. Random sample of 5 items from study in Section 4: original Google translation (GT), results of targeted paraphrasing translation process (TP), and a human reference translation.

However, two issues merited further investigation. First, in the full version of the study, we failed to obtain substantial gain over Google n-best translation for the larger set of sentences. This is problematic, since one would hope that the gains of the approach would not depend simply on having a larger number of hypotheses available to consider, but rather that they would reflect the specific value of generating these alternatives via targeted paraphrase. Second, and related, was the question of why our overall gain in quality was not even larger, since typically one would view an oracle evaluation as the upper bound on expected improvements if a process were to be deployed in the real world. Here we provide some additional analysis focused on these questions.

Error Spans	1	2	3	4	5
Number of Sentences	183	206	161	123	101
Orig→Para	0.78	1.7	0.98	3.76	1.63
Nbest→Para	-0.49	0.32	-1.30	1.19	0.013

Fig. 8. Improvements in Chinese-English study, showing results by number of error spans identified in target sentence.

Figure 8 illustrates the problem. If we look in more detail at the results of the full study, we see that even though we had very substantial gains against the baseline (1-best Google Translate output), our performance against the Google n-best oracle was variable. In the table, the original results in Figure 6, namely the deltas of our approach against the 1-best Google Translation baseline, are labeled Orig→Para, and we now explicitly show the deltas in comparison with Google Translation *n*-best output as Nbest→Para. As before, these are grouped into bins by the number of target-side error spans identified for the sentence, that is, the number of potential paraphrases to be done on the source side. As the table shows, comparisons with Google's n-best translation output yields improvements in only two out of the five subsets of sentences, with a significant decrease of −1.3 BLEU in one of the bins.

Based on a consideration of our overall approach, we hypothesized that the potential of our method might be suffering noise in the system, in the form of poor source-side paraphrases. Quality assurance is a well-known problem in the crowdsourcing community, and our task, like any other, is susceptible to both cheating and to people attempting the task in good faith but simply doing a poor job. This hypothesis led us to a useful analysis of paraphrase quality and its effects, followed by initial investigation of an automatic method for mitigating the problem.

To assess paraphrase quality, we conducted a study on Amazon Mechanical Turk (MTurk) in which we had workers judge the source paraphrases that had been collected in the prior experiment, on a scale from 1 to 5, using the same basic definition of adequacy that is used for studying preservation of meaning in machine translation evaluation. Crucially, the judgments were contextual. That is, to evaluate the quality of paraphrase p' for an original p, the worker was shown the original sentence $\alpha p \beta$ and the sentence $\alpha p' \beta$ with the paraphrase substituted into the identical context, and asked the extent to which the latter, as a full sentence, had the same meaning as the former.[10] It is important to emphasize that this form of quality control is another task that requires only monolingual expertise, on the source side.

We collected multiple in-context judgements for every paraphrase.[11] We found that roughly half of all original paraphrases were poor quality, not reflecting the meaning of the reference sentence according to independent judgments (average ratings $\leq$2 on the 5-point scale). While this reflects the risks of working with Mechanical Turk, it also highlights how easy it is to implement human quality control in the service.

The natural question to ask next is what happens when paraphrases are restricted to cases with reasonable quality, that is, what the results would look like if we added human quality control to Mechanical Turk, or, by the same token, how they would look if we used an alternative approach to crowdsourcing in which participants could be

[10]In preliminary exploration, we obtained quite poor results based on evaluation of paraphrase quality that presents p and p' out of context.

[11]Our intent was to collect three for each, but after the fixed time we allotted to the task, that goal was not met in all cases. We believe this does not substantially affect the results, which are based on mean ratings, although it does mean that those means are associated with higher variance in some cases.

Error Spans	1	2	3	4	5
Number of Sentences	340	211	94	49	24
Orig→Para	1.8589	1.5143	2.7328	3.4984	1.6299
Nbest→Para	1.406	-0.459	1.1298	0.9013	-0.7548

Fig. 9. Improvements with filtering based on paraphrase quality.

relied upon to be working in good faith.[12] We therefore did an analysis similar to the preceding, but based only on the subset of sentences containing error spans for which at least one paraphrase was found to be of reasonable quality (mean rating $\geq$3), and throwing away paraphrases of lower quality for those sentences.

Considering this quality-controlled set of sentences, we find that the number of error spans per sentence is drastically reduced, and the number of sentences with more than five paraphrases is essentially negligible. The plurality of sentences now only contain exactly one error span per sentence. Figure 9 shows the results in the quality-controlled case.

As the table shows, we now see much stronger improvements over the Google Translation 1-best baseline, as well as a sharp increase in improvement over Google Translation n-best results. Particularly interesting is the +1.8 BLEU improvement over original for the 340 sentences with only one error span and the +1.4 BLEU improvement over the Google n-best for the same set. Equally important, the approach now also demonstrates meaningful gains over Google Translation, +0.72 BLEU points, even when the comparison is with Google's n-best rather than 1-best output.

These new results directly address both issues raised at the beginning of this section. The analysis demonstrates a much stronger improvement over the real-world Google Translation 1-best baseline, and it also shows that the method based on targeted paraphrase is doing more than simply exploiting selection from a larger number of translation hypotheses.

These results are very encouraging on their own, and lead to an additional follow-up question: is it possible to perform the quality control step automatically, rather than relying on source-side input from monolinguals?

To address the question, we first conducted an analysis to see whether human judgments of quality of paraphrases actually correspond to gains in the TERp score, that is, gains in translation performance, for individual sentences. The results are thoroughly reassuring: we found that the judgments correlate with improvement in performance with overwhelming statistical significance ($p < 1e - 10$). This enabled us to be confident that our Mechanical Turk task of paraphrase evaluation was actually leading to removal of poor paraphrases that were detracting from our performance.

Next, we explored the possibility of making this distinction automatically. To this end, we developed six very simple heuristics that can be evaluated easily for each paraphrase. These are described briefly with reference to Figure 10.

—*partialCopy*. This feature identifies whether the provided paraphrase is just a partial copy of the original error span. In other words, it identifies paraphrases that are substrings of the error span on which they are based. An example can be seen in Figure 10 where the paraphrase is just the first three characters of the 5-character error span.

[12]One key application planned for this work is translation of children's books in the International Children's Digital Library [Hourcade et al. 2003], which has a substantial base of volunteers willing to help with translation tasks. Exploiting this volunteer base more effectively by going beyond bilingual expertise was, in fact, the original inspiration for this line of research. See Section 7.

Feature	Original (error span underlined)	Paraphrase
partialCopy	...等人在欧洲踢过球不说,李玮锋等人也是久...	踢过球
rearrange	...同捕捞区的建议,并详细说明了其理由。	说细详
othercopy	古巴代表团21日抱怨,...,将会影响拳手水平...	响
superset	...定了一名在约旦首都安曼郊区居住的美国...	名在约旦首都安
english	虽然,受害者能够感觉自己已经安全了,...	感到安全-felt safe
sizeDiff=6	土地不被私有化是我心平气和的底线...	和能得以保持
sizeDiff=2	...最有发言权的应该是亲身经历了...	应该
sizeDiff=0.44	...睹的乘客们.却却如刚刚睡醒般,齐心协力...	刚睡醒般
sizeDiff=0.167	...关注到的,都是美好的表面,...	好

Fig. 10. Examples of six heuristic features used to predict TERp improvement based on targeted paraphrasing of Chinese source sentences.

—*rearrange*. This feature indicates whether the paraphrase consists of the same characters as the error span, but in a different order. The example in Figure 10 shows a paraphrase that is just the reverse of the marked error span.
—*othercopy*. This feature is similar to the partialCopy feature, but it identifies paraphrases that are copies of parts of the original sentence not covered by the error span. The example in Figure 10 shows a situation where the paraphrase is a copy of the 5th character to the left of the final ellipsis.
—*superset*. This feature identifies paraphrases that contain the entire error span as well as padding additional characters either before or after it in the original sentence. This is a common form of cheating, for our Mechanical Turk HITs, since people could just copy and paste a segment surrounding the error span. The example in Figure 10 shows a paraphrase that is just the error span itself, plus 4 preceding characters and 2 following characters in the original sentence.
—*english*. This feature identifies sentences where the Chinese paraphrase included English (either by itself, or mixed in with the Chinese). Even though there are situations where this is not necessarily cheating, it is a possible indication that someone has proposed a poor paraphrase. An example is found in Figure 10.
—*sizeDiff*. The "sizeDiff" feature is a continuous variable that represents the ratio of the paraphrase's length to the length of the error span. We conjecture that values diverging significantly from 1.0 represent inaccurate or improper paraphrases. Examples of four situations, ranging from a sixfold expansion to a sixfold reduction, are shown in Figure 10.

We created a linear model (multiple linear regression, implemented in R), predicting the TERp improvement from the six heuristics. Five of the six were significant predictors, with three significant at $p < 0.001$, and all significant predictors correlated in the expected direction (i.e., the presence of one of the binary heuristics, and an increase in the ratio of original to paraphrase, all lead to a decrease in the performance of the sentence with the paraphrase included). The multiple correlation is significant at $p < 1e - 15$.

As a first foray into automatic quality control, we used the linear model in a predictive fashion (on items not used to create the model) to determine whether or not a proposed paraphrase should be considered unacceptable. Thresholding the value predicted by

the linear model at 0.1, we reject 492 paraphrases, with 465 of them being correct positive rejections. In terms of an accuracy/coverage trade-off, this corresponds to 94.5% accuracy on determinations made on 9.8% of the total data. Applied to the data in our experiment, this would yield a roughly 20% automated removal of poor-quality paraphrases from the system with no human involvement. We are optimistic that with more intricate heuristics and a deeper insight into patters of paraphrasing that represent both cheating and potential improvement, even greater automatic gains will be possible.

6. AUTOMATING ERROR SPAN TARGETING

As we noted in Section 2, the targeted paraphrasing translation process defines a set of human-machine combinations that do not require bilingual expertise. Sections 4 and 5 described human identification of mistranslated spans on the target side, human generation of paraphrases for problematic subsentential spans on the source side, and both automatic hypothesis selection and human selection (via fluency ratings, in Section 3). Human and automatic quality control for paraphrasing was explored in Section 5.

In this section, we take a step toward automating the central piece of the process involving human targeting, by replacing human identification of mistranslated spans with a fully automatic method.[13] The idea behind our automatic error identification is straightforward: if the source sentence is translated to the target and then back-translated, a comparison of the result with the original is likely to identify places where the translation process encountered difficulty.[14] Briefly, we automatically translate source F to target E, then back-translate to produce F' in the source language. We compare F and F' using TERp—which, in addition to its use as an evaluation metric, is a form of string-edit distance that identifies various categories of differences between two sentences. When at least two consecutive edits are found, we flag their smallest containing syntactic constituent as a potential source of translation difficulty.[15]

In more detail, we posit that if an area of back-translation F' has many edits relative to original sentence F, then that area probably comes from parts of the target translation that did not represent the desired meaning in F very well. We only consider consecutive edits in certain of the TERp edit categories, specifically, Deletions (D), Insertions (I), and Shifts (S); the two remaining categories, Matches (M) and Paraphrases (P), indicate that the words are identical or that the original meaning was preserved. Furthermore, we assume that while a single D, S, or I edit might be fairly meaningless, a string of at least two of these types of edits is likely to represent a substantive problem in the translation.

In order to identify reasonably meaningful paraphrase units based on potential errors, we rely on a source language constituency parser. Using the parse, we find the smallest constituent of the sentence containing all of the tokens in a particular error string. At times, these constituents can be quite large, even the entire sentence. To weed out these cases, we restrict constituent length to no more than 7 tokens.

For example, given the following, spans in the italicized phrase in F would be identified, based on the TERp alignment and smallest containing constituent, as shown in Figure 11.

[13]This section contains material we originally reported in Buzek et al. [2010].

[14]Exactly the same insight is behind the "source-side pseudo-reference-based feature" employed by Soricut and Echihabi [2010] in their system for predicting the trustworthiness of translations.

[15]It is possible that the difficulty so identified involves back-translation only, not translation in the original direction. If that is the case, then more paraphrasing will be done than necessary, but the quality of the TP process's output should not suffer.

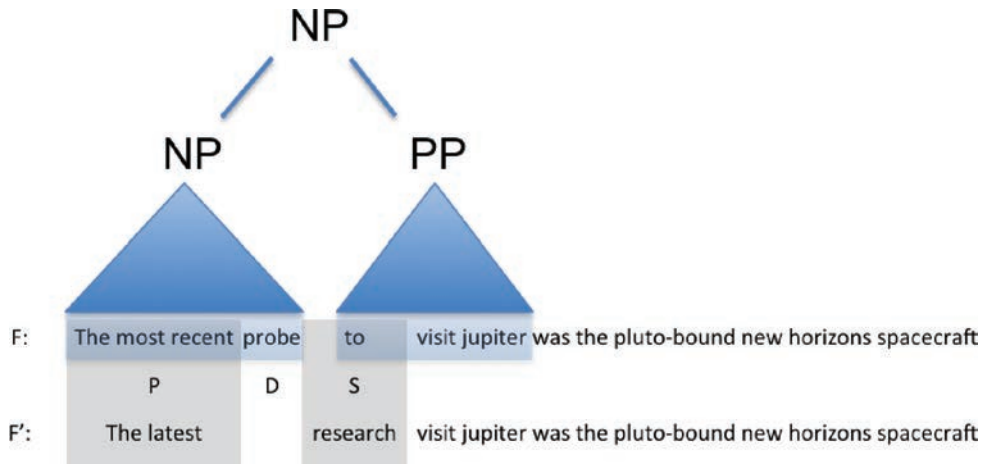

Fig. 11. TERp alignment of a source sentence, F, and its back-translation, F', in order to identify a problematic source span: the top Noun Phrase (NP) is the smallest syntactic constituent subsuming adjacent edits (here, D and S). The figure also illustrates TERp correctly identifying that one of the differences between the two sentences is a paraphrase, which preserves meaning and is therefore not an error.

F: *The most recent probe to visit Jupiter* was the Pluto-bound New Horizons spacecraft in late February 2007.
E: La investigación más reciente fue la visita de Júpiter a Plutón de la envolvente sonda New Horizons a fines de febrero de 2007.
F': The latest research visit Jupiter was the Pluto-bound New Horizons spacecraft in late February 2007.

In order to evaluate this approach, we again use NIST MT08 data, this time going in the English-to-Chinese direction since we are assuming source language resources not currently available for Chinese.[16] We used English reference 0 as the source sentence, and the original Chinese sentence as the target.[17]

The dataset contains 1,357 sentence pairs. Using the earlier described algorithm to automatically identify possible problem areas in the translation, with the Google Translate API providing both the translation and back-translation, we generated 1,780 potential error spans in 1,006 of the sentences, and, continuing the targeted paraphrasing process, we obtained up to three source paraphrases per span, for the problemantic spans in 1,000 of those sentences. (For six sentences, no paraphrases were suggested for any of the problematic spans.) These yielded full-sentence paraphrase alternatives for the 1,000 sentences, which we again evaluated via an oracle study.

For this study we used the TER metric [Snover et al. 2006] rather than TERp. Comparing with the GT output, we find that TP yields a better-translated paraphrase sentence in 313 of the 1000 cases, or 31.3%, and for those 313 cases, TER for the oracle-best paraphrase alternative improves on the TER for the original sentence by 12.16 TER points. Also taking into account the cases where there is no improvement over the baseline, the average TER score improves by 3.8 points. The cost for human

[16]The Stanford parser [Klein and Manning 2002], which we use to identify source syntactic constituents, exists for both English and Chinese, but TERp uses English resources such as WordNet in order to capture acceptable variants of expression for the same meaning. Matt Snover (personal communication) is working on extension of TERp to other languages.
[17]We chose reference 0 because on inspection these references seemed most reflective of native English grammar and usage.

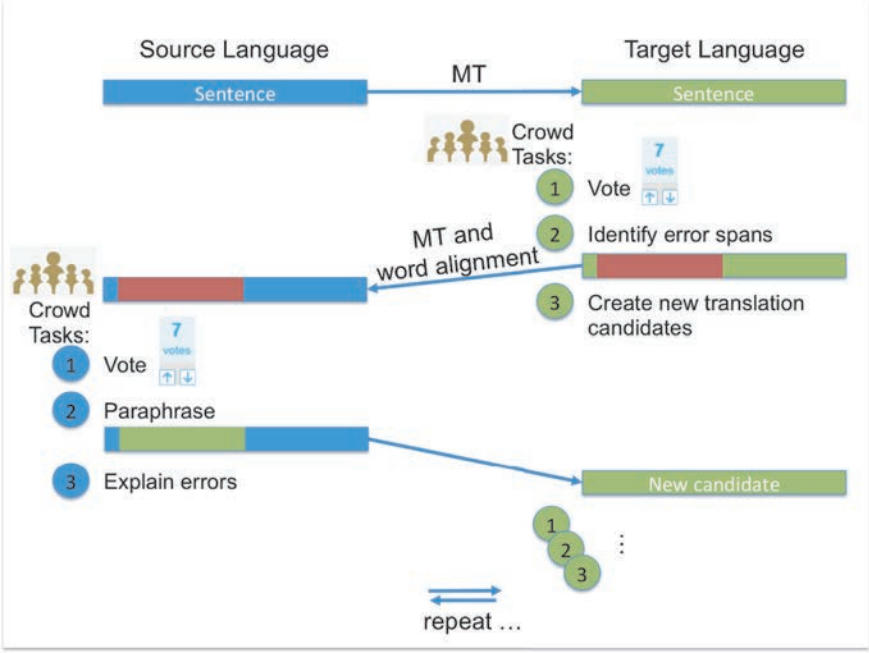

Fig. 12. Collaborative protocol for monolingual translation.

tasks in this study—just paraphrases, since identifying problematic spans was done automatically—was $117.48, or a bit under $0.12 per sentence.

7. BROADER CONTEXT: A MONOLINGUAL PROTOCOL FOR COLLABORATIVE TRANSLATION

In this article, we introduce and evaluate the idea of targeted paraphrasing in isolation. However, as noted in the Introduction, it is also consistent with a context for crowd-sourced translation in which error targeting and paraphrasing are parts of a broader collaborative protocol involving crowds of effectively monolingual users on the source and target side. In Hu et al. [2011], we report on experimentation using a collaborative interface that permits a wider variety of monolingual tasks. Here we briefly summarize that approach, since it provides a broader context for the work, as well as the key results presented there.

The protocol presented in Hu et al. [2011] is illustrated schematically in Figure 12. Monolingual tasks on the target side include not only error span detection, as in this article, but also voting for preferred hypotheses (i.e., human hypothesis selection) and manual creation of new translation candidates (in the same spirit as post-editing). On the source side, monolingual tasks include not only subsentential paraphrasing based on targeted error spans, as in this article, but also voting (based on back-translation of target hypotheses) and "explaining" error spans by manually annotating them, for example, with images (via Google image search) and URLs (for example, links into Wikipedia).

This protocol was tested using translation of children's books between Spanish and German as the task—a specific instance of the problem that originally motivated this line of research, namely the real-world need for a cost-effective way to translate literally thousands of books in the International Children's Digital Library across more than 50 languages [Hourcade et al. 2003].

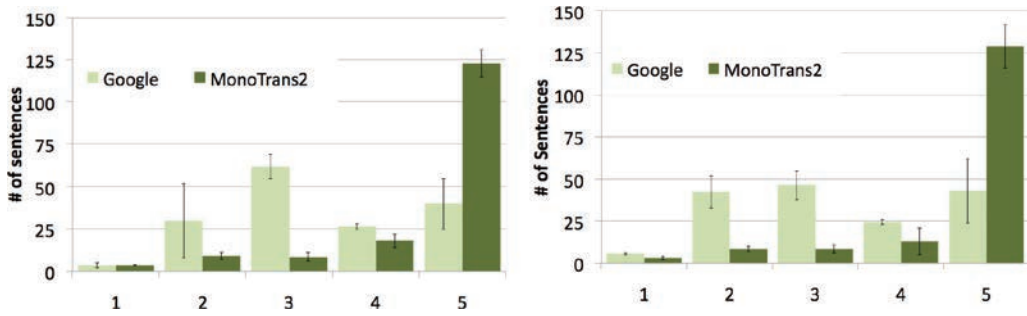

Fig. 13. Results of collaborative protocol for monolingual translation on children's books, showing substantial improvements in fluency (left) and adequacy (right) as judged by bilingual evaluators.

Figure 13 illustrates dramatic improvements in fluency and adequacy (as judged by bilingual evaluators) for the output of the collaborative protocol. As a single figure of merit, the original output of Google Translate produced correct results for only 10% of 162 sentences, aggregating across several books, and this number improved to 68% after the collaborative protocol. "Correct" was defined quite conservatively here, as ratings of 5 for both fluency and adequacy by both of two independent bilingual evaluators.

8. CONCLUSIONS AND FUTURE WORK

In this article we have focused on a relatively less-explored space on the spectrum between high-quality and low-cost translation: sharing the burden of the translation task among a fully automatic system and monolingual human participants, without requiring human bilingual expertise. The monolingual participants in our framework perform straightforward tasks: they identify parts of sentences in their language that seem to have errors, they provide subsentential paraphrases in context, and they judge the fluency of sentences they are presented with (or, in a variant still to be explored, they simply select which target sentence they like the best). Unlike other proposals for exploiting monolingual speakers in human-machine collaborative translation, the human steps here are amenable to automation: in addition to evaluating a mostly human variant of our targeted paraphrasing translation framework, we also assessed a version in which the identification of mistranslated spans to be paraphrased is done automatically. Our experimentation yielded a consistent pattern of results, supporting, via several different measures, the conclusion that targeted paraphrasing can lead to significant improvements in translation.

These initial studies leave considerable room for future work. One important step will be to better characterize the relationship between cost and quality in quantitative terms: how much does it cost to obtain how much quality improvement, and how does that compare with typical professional translation costs? Since our experimentation thus far has been done at a small scale and without "production quality" software, we believe that the approaches we are pursuing could certainly be done both faster and less expensively. But even so, our current results are promising: average cost for the human tasks in our experimentation was under $0.30 per sentence in the test set, and in our experience, professional bilingual translators typically charge $0.15 to $0.25 per word. We did not measure per-sentence translation latency, since our experimentation was based on batch translation, but translating a 1,357-sentence test set took two to three days in an experiment where time was not emphasized and where the real-world setting permits nearly arbitrary degrees of parallelism. This can be compared with a typical turnaround of at least one to two days when hiring professional translators

for even modestly sized jobs. Exploring these questions at scale will involve a complicated ecosystem of workers and cheaters, tasks, and motivations and incentives [Quinn and Bederson 2011]. Zaidan and Callison-Burch [2011] provide detailed discussion of cost and quality with a focus on bilingual translation crowdsourcing, and we have recently begun a collaboration with Callison-Burch to explore issues of monolingual and bilingual translation crowdsourcing within a unified framework.

A related crowdsourcing issue requiring further study is the availability of monolingual human participants for a range of language pairs, in order to validate the argument that drawing on monolingual human participation will significantly reduce the severity of the availability bottleneck. And, of course, in the upper bound in Table V makes quite clear the crucial value added by bilingual translators, when they are available; we hope to explore whether the targeted paraphrasing translation pipeline can improve the productivity of post-editing by bilinguals, making it easier to move toward the upper bound in a cost-effective way.

Another set of issues concerns the underlying translation technology. The value of the approach taken here is likely to vary depending upon the quality of the underlying translation system, and the approach may break down at the extrema, when the baseline translation is either already very good or completely awful. We chose to use Google Translate for its wide availability and the fact that it represents a state-of-the-art baseline to beat; however, in future work we plan to substitute a statistical MT system to which we have developer-level access, such as *cdec* [Dyer et al. 2010], which will permit us to experiment across a range of translation model and language model LM training set sizes, and therefore to vary quality while keeping other system details constant.

More directly connected to research in machine translation, this framework provides a variety of opportunities for advancing the state-of-the-art by combining human and machine components in flexible ways. As one example, the human feedback we are obtaining can provide information about the kinds and distribution of errors in the machine translation system's output; as another, a statistical analysis of manually annotated mistranslation spans could help to identify source-side properties of input spans that are likely mistranslated. Errors could be analyzed according to a taxonomy of translation error types like the one introduced by Vilar et al. [2006], which might contribute valuable data for automatic detection of mistranslations.[18] Our framework also makes it possible to compare human paraphrases with those obtained by automatic methods (e.g., Bannard and Callison-Burch [2005], Callison-Burch et al. [2006], Callison-Burch [2008], and Marton et al. [2009]) on a potentially large scale, which may help improve both our own collaborative translation process and also the state-of-the-art in automatic paraphrasing. More generally, any component in Figure 2 that is represented by a rectangle can be a task for either humans or machines, which means that any such component can serve both as a source of data for evaluation and development of automated methods and as a testbed for those methods. This leads quite naturally to a fully automated pipeline, using algorithms for error span detection (e.g., Section 6), automatic source-side paraphrasing (e.g., Bannard and Callison-Burch and other references cited before), and translation of targeted paraphrase lattices (e.g., Max [2010] and Du et al. [2010]). We plan to implement a fully automatic pipeline of this kind. Finally, we intend to explore the application of our approach in scenarios involving less-common languages, by using a more common language as a pivot or bridge [Habash and Hu 2009].

[18]Thanks to an anonymous reviewer for discussion and suggestions on these points.

ACKNOWLEDGMENTS

The authors would like to thank Chris Callison-Burch and Chris Dyer for their helpful comments and discussion. We would also like to thank the three anonymous EMNLP reviewers of Resnik et al. [2010], and particularly the three anonymous ACM TIST reviewers of this article, for their care in reviewing and their invaluable comments and suggestions.

REFERENCES

99Translations. 2011. 99translations. http://99translations.com/.

Albrecht, J. S., Hwa, R., and Marai, G. E. 2009. Correcting automatic translations through collaborations between mt and monolingual target-language users. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL'09)*. Association for Computational Linguistics, 60–68.

Bannard, C. and Callison-Burch, C. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL'05)*. Association for Computational Linguistics, 597–604.

Bederson, B. B., Hu, C., and Resnik, P. 2010. Translation by iterative collaboration between monolingual users. In *Proceedings of the Conference on Graphics Interface (GI'10)*. 39–46.

Bowker, L. and Barlow, M. 2004. Bilingual concor dancers and translation memories: A comparative evaluation. In *Proceedings of the 2nd International Workshop on Language Resources for Translation Work, Research and Training (LRTWRT'04)*. Association for Computational Linguistics, 70–79.

Buzek, O., Resnik, P., and Bederson, B. 2010. Error driven paraphrase annotation using mechanical turk. In *Proceedings of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, 217–221.

Callison-Burch, C. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'08)*. ACL, 196–205.

Callison-Burch, C. 2009. Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 286–295.

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

Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M., and Zaidan, O. F. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the 5th Joint Workshop on Statistical Machine Translation and MetricsMATR (WMT'10)*. Association for Computational Linguistics, 17–53.

Callison-Burch, C., Koehn, P., and Osborne, M. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL'06)*.

Dabbadie, M., Hartley, A., King, M., Keithj, W., Belis, A. P., Reeder, F., and Vanni, M. 2002. A hands-on study of the reliability and coherence of evaluation metrics. In *Proceedings of the Workshop on Machine Translation Evaluation: Human Evaluators Meet Automated Metrics*. 8–16.

Doyon, J., Taylor, K., and White, J. 1998. The DARPA machine translation evaluation methodology. In *Proceedings of the Annual Meeting of the Association for Machine Translation in the Americas (AMTA'98)*.

Du, J., Jiang, J., and Way, A. 2010. Facilitating translation using source language paraphrase lattices. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Dyer, C. 2007. Noisier channel translation: Translation from morphologically complex languages. In *Proceedings of the 2nd Workshop on Statistical Machine Translation*.

Dyer, C., Lopez, A., Ganitkevitch, J., Weese, J., Ture, F., Blunsom, P., Setiawan, H., Eidelman, V., and Resnik, P. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of the ACL System Demonstrations (ACLDemos'10)*. 7–12.

Dyer, C., Muresan, S., and Resnik, P. 2008. Generalizing word lattice translation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT'08)*.

Dyer, C. and Resnik, P. 2010. Forest translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'10)*.

Esteban, J., Lorenzo, J., Valderr'abanos, A. S., and Lapalme, G. 2004. Transtype2 - An innovative computer-assisted translation system. In *The Companion Volume to the Proceedings of the 42nd Annual Meeting*

of the Association for Computational Linguistics. Association for Computational Linguistics, 94–97. TT2.

FACEBOOK. 2011. Facebook translations app. http://www.facebook.com/translations/.

GETLOCALIZATION. 2011. getlocalization. http://getlocalization.com/.

GUERRA, L. 2003. Human translation versus machine translation and full post-editing of raw machine translation output. M.S. thesis, Dublin City University.

HABASH, N. AND HU, J. 2009. Improving arabic-chinese statistical machine translation using english as pivot language. In *Proceedings of the 4th Workshop on Statistical Machine Translation (StatMT'09)*. Association for Computational Linguistics, 173–181.

HESTER, V., SHAW, A., AND BIEWALD, L. 2010. Scalable crisis relief: Crowdsourced sms translation and categorization with mission 4636. In *Proceedings of the 1st ACM Symposium on Computing for Development (ACM DEV'10)*. ACM Press, New York, 15:1–15:7.

HOURCADE, J. P., BEDERSON, B. B., DRUIN, A., ROSE, A., FARBER, A., AND TAKAYAMA, Y. 2003. The international children's digital library: Viewing digital books online. *Interact. Comput. 15*, 151–167.

HU, C., BEDERSON, B., AND RESNIK, P. 2011. Monotrans2: A new human computation system to support monolingual translation. In *Human Factors in Computing Systems (CHI'11)*. ACM Press, New York.

HUBERDEAU, L.-P., PAQUET, S., AND D' ESILETS, A. 2008. The cross-lingual wiki engine: Enabling collaboration across language barriers. In *Proceedings of the 4th International Symposium on Wikis (WikiSym'08)*. ACM Press, New York, 13:1–13:14.

HWA, R., RESNIK, P., WEINBERG, A., CABEZAS, C., AND KOLAK, O. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Nat. Lang. Eng. 11*, 3, 311–325.

KHADIVI, S., ZENS, R., AND NEY, H. 2006. Integration of speech to computer-assisted translation using finite-state automata. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics Main Conference Poster Sessions (COLING/ACL'06)*. Association for Computational Linguistics, 467–474.

KLEIN, D. AND MANNING, C. D. 2002. Fast exact inference with a factored model for natural language parsing. In *Proceedings of the Conference on Advances in Neural Information Processing Systems 15 (NIPS'02)*. S. Becker, S. Thrun, and K. Obermayer, Eds., MIT Press, 3–10.

KOEHN, P. 2009. A web-based interactive computer aided translation tool. In *Proceedings of the ACL-IJCNLP Software Demonstrations*. Association for Computational Linguistics, 17–20.

LAURIAN, A.-M. 1984. Machine translation: What type of post-editing on what type of documents for what type of users. In *10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*.

MARTON, Y., CALLISON-BURCH, C., AND RESNIK, P. 2009. Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 381–390.

MAX, A. 2009. Sub-sentencial paraphrasing by contextual pivot translation. In *Proceedings of the Workshop on Applied Textual Inference*. Association for Computational Linguistics, 18–26.

MAX, A. 2010. Example-based paraphrasing for improved phrase-based statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Cambridge, MA.

MEEDAN. 2011. Meedan: An arabic-english forum using machine translation with expert corrections. http://news.meedan.net/.

MORITA, D. AND ISHIDA, T. 2009. Designing protocols for collaborative translation. In *Proceedings of the 12th International Conference on Principles of Practice in Multi-Agent Systems (PRIMA'09)*. Springer, 17–32.

MUNRO, R. 2010. Haiti emergency response: The power of crowdsourcing and sms. Relief 2.0 in haiti, Stanford, CA.

OCH, F. J., GILDEA, D., KHUDANPUR, S., SARKAR, A., YAMADA, K., FRASER, A., KUMAR, S., SHEN, L., SMITH, D., ENG, K., JAIN, V., JIN, Z., AND RADEV, D. R. 2004. A smorgasbord of features for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'04)*. 161–168.

PAPINENI, K., ROUKOS, S., WARD, T., AND ZHU, W.-J. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL'02)*. Association for Computational Linguistics, 311–318.

PRZYBOCKI, M., PETERSON, K., BRONSART, S., AND SANDERS, G. 2009. The nist 2008 metrics for machine translation challenge–Overview, methodology, metrics, and results. *Mach. Trans. 23*, 71–103.

QUINN, A. J. AND BEDERSON, B. B. 2011. Human computation: A survey and taxonomy of a growing field. In *Human Factors in Computing Systems (CHI'11)*. ACM.

Resnik, P., Buzek, O., Hu, C., Kronrod, Y., Quinn, A. J., and Bederson, B. B. 2010. Improving translation via targeted paraphrasing. In *Proceedings of the Conference on Empirical Methods in Natural Language (EMNLP'10)*. ACL, 127–137.

Shahaf, D. and Horvitz, E. 2010. Generalized task markets for human and machine computation. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence.*

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the Association for Machine Translation in the Americas*. 223–231.

Snover, M., Madnani, N., Dorr, B., and Schwartz, R. 2009. TER-Plus: Paraphrases, semantic, and alignment enhancements to translation edit rate. *Mach. Trans. 23*, 2–3, 117–127.

Soricut, R. and Echihabi, A. 2010. Trustrank: Inducing trust in automatic translations via ranking. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 612–621.

Ted. 2011. Ted: Ideas worth spreading. http://www.ted.com/translate/.

Twitter. 2011. Twitter translation center. http://translate.twttr.com/welcome.

Vilar, D., Xu, J., D'haro, L. F., and Ney, H. 2006. Error analysis of machine translation output. In *Proceedings of the International Conference on Language Resources and Evaluation*. 697–702.

Yarowsky, D., Ngai, G., and Wicentowski, R. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the 1st International Conference on Human Language Technology Research (HLT'01).* Association for Computational Linguistics, 1–8.

Zaidan, O. and Callison-Burch, C. 2011. Crowdsourcing translation: Professional quality from nonprofessionals. In *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT'11)*. ACL.

Received March 2011; revised July 2011; accepted November 2011