Improving Translation via Targeted Paraphrasing

Philip Resnik
Linguistics and UMIACS
University of Maryland
resnik@umd.edu

Olivia Buzek
Linguistics and Computer Science
University of Maryland
olivia.buzek@gmail.com

Chang Hu
Computer Science
University of Maryland
changhu@cs.umd.edu

Yakov Kronrod
Linguistics and UMIACS
University of Maryland
yakov@umd.edu

Alex Quinn
Computer Science
University of Maryland
aq@cs.umd.edu

Benjamin B. Bederson
Computer Science and UMIACS
University of Maryland
bederson@cs.umd.edu

Abstract

Targeted paraphrasing is a new approach to the problem of obtaining cost-effective, reasonable quality translation that 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. Evaluations demonstrate that this approach can yield substantial improvements in translation quality.

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 non-professional translators, which can be remarkably effective (Munro, 2010). However, all these alternatives face a central availability bottleneck: they require the participation of humans with bilingual expertise.

In this paper, 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*, i.e. 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.

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). 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. 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.

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 127–137,
MIT, Massachusetts, USA, 9-11 October 2010.

Bederson et al. (2010) describe a similar protocol in which cross-language communication is enhanced by metalinguistic communication in the user interface. 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.

We call the technique used here *targeted paraphrasing*. 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. This technique can be viewed as compatible with the richer protocol- and game-based approaches, but it is considerably simpler; in Sections 2 through 4 we describe the method and present evaluation results on Chinese-English translation. Unlike other approaches, the technique also offers clear opportunities to replace human participation with machine components if the latter are up to the task; we discuss this in Section 5 before wrapping up in Section 6 with conclusions and directions for future work.

2 Targeted Paraphrasing

The starting point for our approach 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; Dyer and Resnik, 2010). Here we apply it 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 pourvoir le siége au Sénat de Sen. Ted Kennedy, qui est décédé récemment.

Operationally, then, translation with targeted paraphrasing includes the following steps.

Initial machine translation. For this paper, 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 above, 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 5, we describe a heuristic but effective method for performing this fully automatically. Du et al. (2010), in this proceedings, 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 paper, 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), in these proceedings, 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 non-overlapping 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.

Machine translation of alternative sentences. The alternative source sentences, produced via paraphrase, are sent through the same MT system, and a single-best translation hypothesis is selected, e.g. 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); cf. Du et al. (2010). One could 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 paper.

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, e.g. "(H_20)". The shortest sentence in this set has 12 Chinese characters, the longest has 54.[1]

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).[2] The Chinese source span paraphrases were then used to construct full-sentence paraphrases, which were retranslated, once again by Google Translate, to produce the output of the targeted paraphrasing translation process.

[1] Note that this page is *not* a translation of the corresponding English Wikipedia page or vice versa.

[2] 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.

The initial translation outputs from Google Translate (GT) and the results of the targeted paraphrasing translation process (TP) were evaluated according to widely used critera of fluency and adequacy. 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. Those are respective gains of 0.82 and 1.21 over the baseline initial MT output, each on a 5-point scale.

Figure 1 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

As a followup to our pilot study, we conducted an evaluation using Chinese-English test data taken from the NIST MT'08 machine translation evaluation, in order to obtain fully automatic translation evaluation scores. We report on results for 49 sentences of the 1,357 in this data set. 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.

Target English speakers identified 115 potential mistranslation spans, or 2.3 spans per sentence, that yielded at least one source paraphrase on the source Chinese side. Chinese speakers provided 138 valid paraphrases. The entire cost for the human tasks in this experiment was $5.06, or a bit under $0.11 per sentence on average.[3]

Table 1 reports on the results, evaluating in standard fashion using BLEU with the four English MT'08 references for each Chinese sentence. Since the targeted paraphrasing translation process (TP) produces multiple hypotheses — one automatic translation output per sentential paraphrases — we selected the single best output for each sentence by

[3] Invalid paraphrase responses were rejected, i.e. zero-cost.

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.

Figure 1: 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.

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

Table 1: Results on a 49-sentence subset of the NIST MT'08 Chinese-English test set

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 yielded an improvement of 1.68 BLEU points on the 49-sentence test set (TP one-best).

One could argue that this result 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 evaluate each hypothesis against the reference translations.[4] The resulting BLEU score for the full set showed negligible improvement (GT n-best oracle).

We did a similar oracle-best calculation using TERp for targeted paraphrasing (TP oracle). The result shows a potential gain of 2.46 BLEU points over the baseline, if the best scoring alternative from the targeted paraphrasing process were always chosen.

In addition to aggregate scoring using BLEU, we also looked at oracle results on a per-sentence basis using TERp (since BLEU more appropriate to use at the document level, not the 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.[5] A fairer measure is the average obtained when scoring zero gain for the 17 sentences where no improvement was obtained; taking these into account, i.e. assuming an oracle who chooses the original translation if none of the paraphrase-based alternatives are better, the average improvement over the entire set of 49 sentences is 5.46 TERp points.

Although we have obtained results on only a small subset of the full NIST MT'08 test set, our automatic evaluation confirms the qualitative impressions in Figure 1 and the subjective ratings results obtained in our pilot study in Section 3. The TP oracle results establish that by taking advantage of monolingual human speakers, it is possible to obtain quite substantial gains in translation quality. 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.

The last line in Table 1 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 were then averaged. The value of this upper bound is quite consistent with the bound computed similarly by Callison-Burch (2009).

5 English-Chinese Evaluation

As we noted in Section 2, the targeted paraphrasing translation process defines a set of human-machine combinations that do not require bilingual expertise. The previous section described human identification of mistranslated spans on the target side, human generation of paraphrases for problematic sub-sentential spans on the source side, and both automatic hypothesis selection and human selection (via fluency ratings, in Section 3).

In this section, we take a step toward more automated processing, replacing human identification of mistranslated spans with an a fully automatic method.[6] The idea behind our automatic error identification is straightforward: if the source sentence

[4] 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).

[5] "Gains" refer to a lower score: since TERp is an error measure, lower is better.

[6] This section contains material we originally reported in Buzek et al. (2010).

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 .

Figure 2: 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.

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.[7] 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.[8]

In more detail, we posit that if an area of backtranslation 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 those 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

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.

spans in the the bolded phrase in F would be identified, based on the TERp alignment and smallest containing constituent as shown in Figure 3.

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.[9] We used English reference 0 as the source sentence, and the original Chinese sentence as the target.[10]

The data set comprises 1,357 sentence pairs. Using the above 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 weres 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 is available 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 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] Exactly the same insight is behind the "source-side pseudo-referencebased feature" employed by Soricut and Echihabi (2010) in their system for predicting the trustworthiness of translations.

[8] 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.

[9] 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.

[10] We chose reference 0 because on inspection these references seemed most reflective of native English grammar and usage.

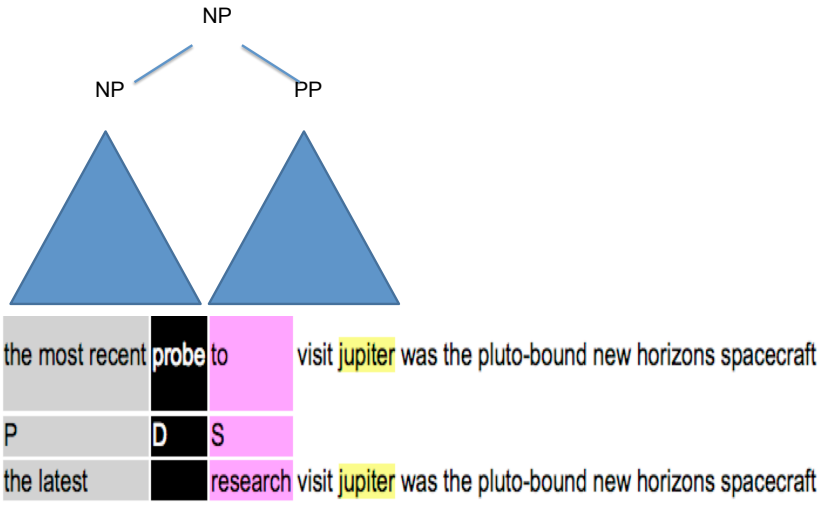

Figure 3: TERp alignment of a source sentence and its back-translation in order to identify a problematic source span.

6 Conclusions and Future Work

In this paper 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 this framework perform straightforward tasks: they identify parts of sentences in their language that seem to have errors, they provide sub-sentential 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, and 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 the conclusion that targeted paraphrasing can lead to significant improvements in translation, via several different measures. First, a very small pilot study for Chinese-English translation in Wikipedia provided preliminary validation that translation fluency and accuracy can be improved quite significantly for a set of fairly chosen test sentences, according to human ratings. Second, a small experiment in Chinese-English translation using standard NIST test sentences suggested the potential for dramatic gains using the BLEU and TERp scores, with oracle improvements of 2.46 points and 5.46 points, respectively. In addition, a non-oracle experiment, selecting the best hypothesis according to the MT system's model score, yielded a gain of nearly 1.7 BLEU points. And third, in a large scale evaluation of the approach using English-Chinese translation of 1,000 sentences, this time automating the step of identifying potentially mistranslated parts of source sentences, the oracle results demonstrated that a gain of nearly 4 TER points is available.

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 of $0.25 per word? This question is closely connected with the dynamics of crowdsourcing platforms such as Mechanical Turk — the cost *per sentence* in these experiments works out to be around $0.12, but translation on a large scale will involve a complicated ecosystem of workers and cheaters, tasks and motivations and incentives (Quinn and Bederson, 2009). 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 1 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. A reviewer correctly notes that 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 our own statistical MT systems, 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 improving fully automatic statistical MT systems. We plan to implement a fully automatic targeted paraphrasing translation pipeline, using the automated methods discussed when introducing the pipeline in Section 2, including translation of targeted paraphrase lattices (cf. (Max, 2010; Du et al., 2010)). 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).

Acknowledgments

This work has been supported in part by the National Science Foundation under awards BCS0941455 and IIS0838801. The authors would like to thank three anonymous reviewers for their helpful comments, and Chris Callison-Burch and Chris Dyer for their helpful comments and discussion.

References

Joshua S. Albrecht, Rebecca Hwa, and G. Elisabeta Marai. 2009. Correcting automatic translations through collaborations between mt and monolingual target-language users. In *EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 60–68, Morristown, NJ, USA. Association for Computational Linguistics.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 597–604, Morristown, NJ, USA. Association for Computational Linguistics.

Benjamin B. Bederson, Chang Hu, and Philip Resnik. 2010. Translation by iterative collaboration between monolingual users. In *Graphics Interface (GI) conference*.

Lynne Bowker and Michael Barlow. 2004. Bilingual concordancers and translation memories: a comparative evaluation. In *LRTWRT '04: Proceedings of the Second International Workshop on Language Resources for Translation Work, Research and Training*, pages 70–79, Morristown, NJ, USA. Association for Computational Linguistics.

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

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

Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 286–295, Singapore, August. Association for Computational Linguistics.

Jinhua Du, Jie Jiang, and Andy Way. 2010. Facilitating translation using source language paraphrase lattices.

In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Cambridge, MA, October. Association for Computational Linguistics.

Chris Dyer and Philip Resnik. 2010. Forest translation. In *NAACL'10*.

C. Dyer, S. Muresan, and P. Resnik. 2008. Generalizing word lattice translation. In *Proceedings of HLT-ACL*, Columbus, OH.

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

José Esteban, José Lorenzo, Antonio S. Valderrábanos, and Guy Lapalme. 2004. Transtype2 - an innovative computer-assisted translation system. In *The Companion Volume to the Proceedings of 42st Annual Meeting of the Association for Computational Linguistics*, pages 94–97, Barcelona, Spain, jul. Association for Computational Linguistics. TT2.

Nizar Habash and Jun Hu. 2009. Improving arabic-chinese statistical machine translation using english as pivot language. In *StatMT '09: Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 173–181, Morristown, NJ, USA. Association for Computational Linguistics.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Nat. Lang. Eng.*, 11(3):311–325.

Shahram Khadivi, Richard Zens, and Hermann Ney. 2006. Integration of speech to computer-assisted translation using finite-state automata. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 467–474, Morristown, NJ, USA. Association for Computational Linguistics.

Dan Klein and Christopher D. Manning. 2002. Fast exact inference with a factored model for natural language parsing. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, *Advances in Neural Information Processing Systems 15 - Neural Information Processing Systems, NIPS 2002*, pages 3–10. MIT Press.

Philipp Koehn. 2009. A web-based interactive computer aided translation tool. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pages 17–20, Suntec, Singapore, August. Association for Computational Linguistics.

Anne-Marie Laurian. 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*.

Aurélien Max. 2009. Sub-sentencial paraphrasing by contextual pivot translation. In *Proceedings of the 2009 Workshop on Applied Textual Inference*, pages 18–26, Suntec, Singapore, August. Association for Computational Linguistics.

Aurélien Max. 2010. Example-based paraphrasing for improved phrase-based statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Cambridge, MA, October. 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. Haiti emergency response: the power of crowdsourcing and SMS. Relief 2.0 in Haiti, Stanford, CA.

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alexander Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir R. Radev. 2004. A smorgasbord of features for statistical machine translation. In *HLT-NAACL*, pages 161–168.

Alex Quinn and Benjamin B. Bederson. 2009. A taxonomy of distributed human computation. Technical Report HCIL-2009-23, University of Maryland, October.

D. Shahaf and E. Horvitz. 2010. Generalized task markets for human and machine computation. In *AAAI 2010*, July.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Matt Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. TER-Plus: Paraphrases, Semantic, and Alignment Enhancements to Translation Edit Rate. *Machine Translation*.

Radu Soricut and Abdessamad Echihabi. 2010. Trustrank: Inducing trust in automatic translations via ranking. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 612–621, Uppsala, Sweden, July. Association for Computational Linguistics.

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.