Translation by Iterative Collaboration between Monolingual Users

Benjamin B. Bederson[1,2], Chang Hu[1,2] and Philip Resnik[1,3,4]

[1]Computer Science Department, [2]Human-Computer Interaction Lab,
[3] Institute for Advanced Computer Studies, [4] Department of Linguistics,
University of Maryland

ABSTRACT

In this paper we describe a new iterative translation process designed to leverage the massive number of online users who have minimal or no bilingual skill. The iterative process is supported by combining existing machine translation methods with monolingual human speakers. We have built a Web-based prototype that is capable of yielding high quality translations at much lower cost than traditional professional translators. Preliminary evaluation results of this prototype confirm the validity of the approach.

KEYWORDS: Monolingual, translation, translation interface, human computation, distributed human computation, wisdom of crowds, crowdsourcing, machine translation.

INDEX TERMS: H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous

1 INTRODUCTION

An enormous potential exists for solving certain classes of computational problems through rich collaboration between humans and computers. Humans alone are expensive and can be surprisingly slow. In our recent effort to hire a commercial service to translate just a few pages of text from English to Mongolian, the task was estimated to require over a month. Translation between two uncommon languages is even more problematic (e.g. between Mongolian and Hungarian). Despite significant recent advances, machine translation (MT) remains a crucial problem and fully automated high quality translation remains a distant dream for the vast majority of the world's language pairs. Usable translation quality can sometimes be obtained by statistical MT systems, but only for a minority of language pairs, and only in use cases where sufficient training text is available and the material being translated is reasonably similar to the material on which the system was trained.

Using the Web to reach non-professional human translators holds promise, and there has been some initial success with distributing translation over a crowd of bilingual users [1]. However, compared to the total user population, the potential translator population is still small. For example, while Wikipedia currently has about 75,000 active contributors, there are fewer than 800 translators [2].

With a much larger number of potential human helpers who speak only the source or target language, but not both, it seems natural to ask whether some combination of machine translation with volunteer *monolingual* speakers could result in high quality translation. Callison-Burch's exploration of "post-editing" (human correction of MT output) and human guided decoding by monolingual target-language speakers suggests that the idea has potential: he improves on automatic translation by using monolingual human participants as sophisticated language models for the target language [12]. We are taking this a step further, to include interaction with a monolingual source-language speaker, as well.

We propose a rethinking of the translation problem to bring together translation technology and human-computer interaction, producing a framework for translation that will exploit imperfect technology and limited human abilities in tandem to achieve capabilities neither can achieve alone.

The core of this framework is an iterative protocol in which the human participants work together to make sense of machine translations, and introduce redundant information to make their intended meanings clearer. Figure 1 shows an example where a French sentence is first translated by an MT system, which is then refined by a monolingual English speaker making her best guess as to the intended meaning. The result is back-translated to French, and the monolingual French speaker corrects the MT output using her knowledge of the intended meaning, as well as limited understanding of the English speaker's guess. (Notice how the rephrased French sentence contains *nous*, which preserves the meaning "we" introduced by the English guess.) The process is then repeated. At each stage, the users can add redundant information of various kinds to help clarify the meaning. Even with only simple types of redundant information in this (real) example (pictures and highlighted correct/incorrect spans of text), the translation process converges to an accurate translation after four steps.

Why can this protocol work? First, natural language is redundant and humans are good at making inferences, even in the face of highly deformative translation channels. Second, shared context (e.g., existing images, translated Wikipedia pages, multiple discussions of the same event) supports linguistic communication independent of the translation channel. Third, people are capable of learning on the fly from their interactions. In earlier experiments, we showed that people who do not speak the same language can nonetheless work together to improve automated translations [16].

The key idea in our monolingual translation protocol is to take advantage of user interface concepts, external knowledge sources, and the interactive nature of the protocol in order to increase the level of redundancy available to the receiver, shifting problems from more to less problematic error categories.

This protocol makes it possible to detect and correct some translation errors, and to at least identify some passages that have errors even if they are not correctable given the available information. For example, "has cheeseburger" is a detectable

{bederson, changhu}@cs.umd.edu, resnik@umiacs.umd.edu

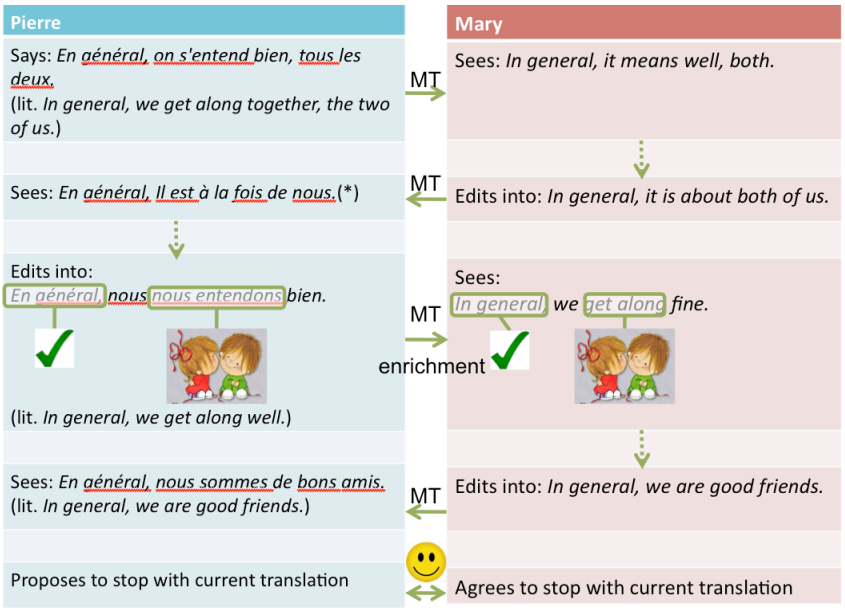

Figure 1. An example of the iterative translation process taken from our experiments. A solid arrow indicates a pass through the system; A dotted arrow indicates an user action. Literal translations of French sentences are shown. The asterisk (*) indicates ungrammatical French MT output. Note the annotations attached to the French phrases are carried over to the English side.

error, even if it is not clear whether the intended meaning was "has cheeseburgers" or "have a cheeseburger". Back-translating a refinement and carrying along redundant information, e.g. a *picture* of multiple cheeseburgers, might help convey which of those alternatives the English speaker guessed, presenting the opportunity for confirmation or further correction.

In the remainder of the paper, we provide some relevant background; expand on the iterative translation protocol in more detail, and present promising results from pilot experimentation.

2 Background

During the last twenty years, a revolution has taken place in computational research on translation: MT systems that used to rely on human knowledge about grammar and meaning provided by language experts have been replaced by systems that learn statistical models from large collections of translated text. This change in approach has made it possible to translate unrestricted input from a far broader spectrum of languages. For example, Google's statistical MT engine today will make a passable attempt at translating among dozens of languages. But "passable" is often far from adequate, and for most language pairs not involving English, we still lack even the most basic ability to create comprehensible translations that preserve basic meaning. Yet many use cases for translation require, if not a guarantee of high quality, then at least the ability to identify the existence of uncorrected errors, in order to inform downstream decision making.

The idea of human-machine collaboration in translation (human-assisted MT or machine-assisted human translation) is a kind of distributed human computation [26]. A number of related approaches have been pursued before. For use cases requiring reliable translation, supportive technologies such as translation memory have existed for years [19], and some researchers have exploited statistical MT modelling to build translation environments that help bilingual human translators do their work more efficiently [15]. With respect to interactive cross-language communication between monolingual users, automatic translation has been deployed in speech-to-speech translation devices [20][25][6] and multilingual chat [21]. Presentation of translation alternatives to monolingual target language readers, rather than a single translation, has been explored in the Linear B system [12] with some promising results. Lemmatic machine translation [31] has also integrated MT with monolingual human editing in both rephrasing the source text (encoding) and inferring the translation (decoding). As its name suggests, it is more focused on the humans helping the machine (to obtain a "passable" translation) rather than on computer-supported iterative collaboration between two humans (for higher quality translation).

Of course, it is possible to use the Web to obtain fully automatic translations (e.g., Google Translate, Babel Fish, etc.) or to request low-cost human translation (e.g., Amazon Mechanical Turk [22]). Recently Meedan.net has introduced a site supporting machine translation with post-editing by a community of volunteer translators, for materials in Arabic and English [10]. Worldwide Lexicon has also introduced a Firefox plug-in with similar functionality [3], and Yeeyan.com is a volunteer-based community translation effort involving translation of Web content, primarily English blogs, into Chinese.

Such previous efforts involving human-machine interactions in translation can be seen as defining a space with two dimensions: bilingual participation and quality. When a bilingual person is involved, the process can yield high or low quality depending on the quality and availability of a suitable translator. Low cost translation on the Web (e.g. soliciting translations via Mechanical Turk) can be quite unreliable. In our own pilot studies, we have found that some Mechanical Turk participants simply use online translation engines rather than performing translations themselves. This is a potential problem with "crowdsourcing" approaches to translation. In the absence of a bilingual participant, real-time interaction between conversational participants (in chat or speech-to-speech settings) can sometimes achieve shared understanding of the intended meaning (e.g. in lemmatic machine translation). Even when achieved, however, shared understanding is no guarantee that a fluent and accurate target sentence is ever actually

		Quality	
		Low	High
Bilingual human?	Yes	Web human translators (e.g. obtained by means of Mechanical Turk)	Conventional and machine-assisted translation (e.g. translation memory, post-editing)
	No	Automatic MT, MT-enabled chat, Speech to speech translation	Least explored

Table 1. Space of Translation Processes

produced. The same observation holds true of pictorial approaches connected with augmented and assisted communication [23][33], including our own group's investigations of picture-based communication among children [18].

Compared to previous research, our focus is on the least explored and most challenging quadrant of the space: achieving high quality translation in the absence of a human participant who is fluent in the source and target languages.

It is worth noting that work on participatory games [4][7] also demonstrates strong potential in engaging the public to help solve hard problems. From image labelling to protein folding, it is clear that it is possible to engage a broad community of people to do meaningful work. These approaches to date, though, focus primarily on using the computer to supply data, communications, and aggregations. One of the new dimensions that we are adding is the use of computers as a more active partner in the computation itself. Our goal is to solve hard problems, and our model is to take advantage of whatever computational resources are available. As the computers get better, people can participate less. And if we have very skilled human participants, they can compensate for weak computational support.

Our work shares some significant elements in common with an interesting and similarly motivated approach proposed independently by Morita and Ishida [24].[1] Like us, they define a translation/back-translation protocol involving monolingual revisions on each side of a machine translation system. Crucially, however, their users are limited to source- or target-side editing, augmented only by requests for full-sentence paraphrase when the machine translation or back translation is incomprehensible. In contrast, we introduce a rich array of user interface concepts, enabling users to focus on regions of interest within the sentence, and to collaborate across the language barrier using language-independent annotations attached to corresponding phrases in both the source and the target texts. Since these annotations are not affected by translation, they can facilitate the establishment of common ground between the source and target language speakers.

3 Translation By Collaboration

Our translation protocol iteratively improves translation quality over a poor translation channel via a form of negotiation between two participants with imbalanced language skills. Figure 2 illustrates the protocol with French and English as the source and target languages, respectively. The figure shows two monolingual participants on either end of an automatic translation channel. The French speaker begins with a source language sentence S_0 to be translated, and this is sent through an automatic translation system that produces an unknown quality English sentence T_1 from S_0. The English speaker edits the automatically generated translation to produce a grammatical English sentence T_2, which is passed back through an automatic system to produce a French sentence S_1. This is corrected to produce grammatical French and the result, S_2, sent back through the channel, translated into T_3, and so forth.

At each step, therefore, the human task is to: a) infer the intended meaning to the extent possible given all the information available; b) correct the text to produce a grammatical sentence in one's own language conveying that meaning; and c) use the interface to provide feedback to the other user. For the source language speaker, who knows the intended meaning, this task is equivalent to translation post-editing [8] and directly analogous to the editing process in MT evaluations involving Human Translation Edit Rate (HTER, [30]), in which human editors make the fewest modifications required to MT system output in order to capture the complete meaning of a reference (i.e. true) translation. In contrast, the target language speaker must infer the intended meaning, much as any user of Google Translate or Babel Fish does when reading imperfect translation from an unfamiliar language; this task is analogous to HTER editing in the absence of a reference translation. The goal state in the present example is where the French speaker, looking at S_i, is sufficiently confident that the originating English sentence T_{i+1} contained the correct meaning. Since T_{i+1} is fluent and grammatically correct (by definition of the English speaker's task), it therefore constitutes a valid translation and the task is complete.

By itself, there would seem to be little reason for confidence that this back-and-forth process should converge on a correct translation. However, as Figure 1 illustrated, this monolingual translation protocol involves more than just translation, editing, and back-translation: it is designed to exploit the fact that increased redundancy leads to more successful communication.

The importance of redundancy in linguistic communication is well established. Redundancy can be characterized as the quantity of information (measured in bits) used in transmitting a message over and above the number of bits in the message itself [29]. Languages contain a variety of phonological, syntactic, semantic, and pragmatic mechanisms that help the listener narrow the hypothesis space for the intended message via redundancy – one common illustration of such constraints involves *rmvng ll th vwls frm th wrds nd shwng tht th rdr cn stll ndrstnd th sntnc*, and noiseless data compression of natural language relies on the fact that linguistic redundancy exists. More generally, we recognize that people manage to communicate successfully in challenging circumstances whether they are in a noisy bar, using a poor quality cell phone connection, playing with a young child, or talking to someone who doesn't speak their language very well. People adapt to all of these situations through a combination of linguistic

[1] It appears that we [16] and Morita and Ishida [24] developed very similar ideas simultaneously.

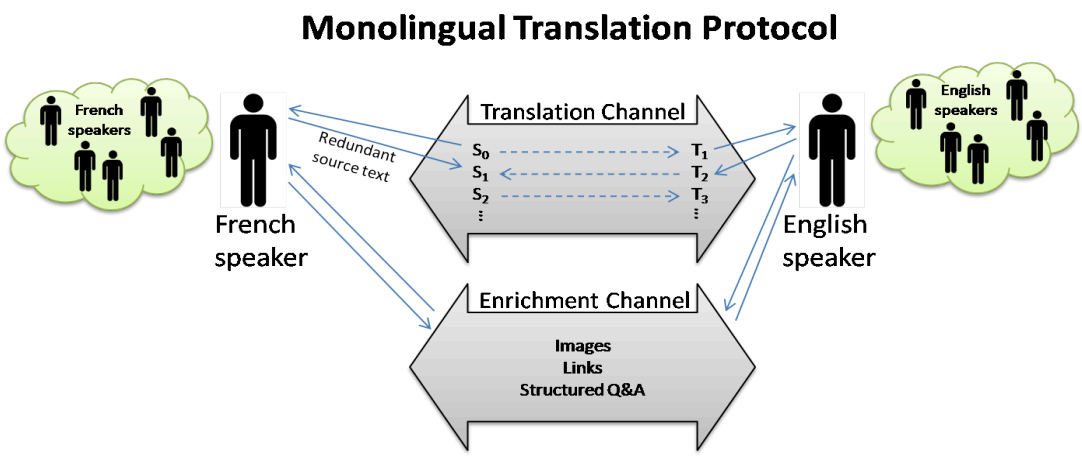

Figure 2. Round-trip protocol. Dashed arrows show machine translation and solid arrows show human editing.

constraints, world knowledge, shared context, and clarification requests.

In the following examples, we classify translation errors into three types. These examples are all in English for ease of exposition, but they are intended to illustrate sentences in the source and target languages.

1. **Errors that are both detectable and correctable**, for example in the target sentence "Everybody has hear story about Cinderella" when the correct source meaning is "Everybody has heard the story about Cinderella". These are often grammatical errors that a monolingual speaker can fix thanks to linguistic redundancy and shared context.
2. **Errors that are detectable but not correctable**, as in "Everybody has heard the business by Cinderella" versus the correct meaning in "Everybody has heard the story about Cinderella". These are errors that a native speaker can identify – clearly "business by Cinderella" is an incorrect translation of *something* – but cannot fix with confidence. (This example is constructed but plausible. The French word *histoire* means "story" but can also mean "business".)
3. **Undetectable errors**, for example "Everybody loves the story about Cinderella" instead of "Everybody has heard the story about Cinderella." In these cases, a fluent and plausible communication gives the receiver no reason to suspect an error has occurred.

By means of user interface design, external knowledge sources, and interaction between participants, our protocol can increase the level of redundancy available to the receiver and shift problems translation errors from more to less problematic error categories. For example, the system will support annotation of the primary message with source language synonyms and sub-sentential paraphrases: if the French speaker believes the concept *histoire* ("story") was misunderstood, he can use the system to indicate that *histoire*, *conte* ("tale"), *récit* ("story"), and *légende* ("legend") are conceptually similar words. Even noisy translations of these words, together with the context, are likely to turn the detectable error "the business by Cinderella" into one that can be corrected. Similarly, by linking images connected with hearing to the verb mistranslated as "love", the substitution of *love* for *hear* can be made detectable and possibly even correctable: A French-English dictionary maps French *entendre* to English *hear*, which produces images in (English) Google image search that are likely to clarify the intended meaning. (In this case, a search could be done directly in French Google image search, as well.)

These examples serve to illustrate the general idea of an enrichment channel accompanying the MT channel, which can be summarized by the following principles. The first, motivated by information theory and discussed earlier, can be called the principle of redundancy: the recoverability of information conveyed over a noisy channel is improved by augmenting the message with redundant information. The second might be called the principle of mutual knowledge: essentially, that successful linguistic communication depends on, and also creates, shared context [13].

All of the above mechanisms for enrichment and feedback raise questions about how to link information connected to part of a sentence in one's own language to the corresponding part of the sentence in the other participant's language. (If a user marks "the story of Cinderella" as having been translated correctly, how do we find the piece of the sentence that phrase was translated from, in order to convey that information back to the other user?) We solve this problem by means of "annotation projection" [17][32], a technique that uses word-level alignments (between corresponding words in the source and target sentences) as a bridge between the two sentences.

Finally, it is also important to question whether mechanisms for adding information and context may also bring in too much additional noise to make it useful. As to this concern, we are encouraged by the results obtained by Callison-Burch in the Linear B system [12]. Monolingual target-language users in that system were presented with all available phrase-level translations accessible by an underlying statistical MT system, and they demonstrated significant ability to pick out intended sub-sentential phrase translations (translations of phrases in a sentence) among a plethora of alternatives by employing their rich target-language knowledge (in effect, serving as a human language model) as well as the full range of their knowledge about the world.

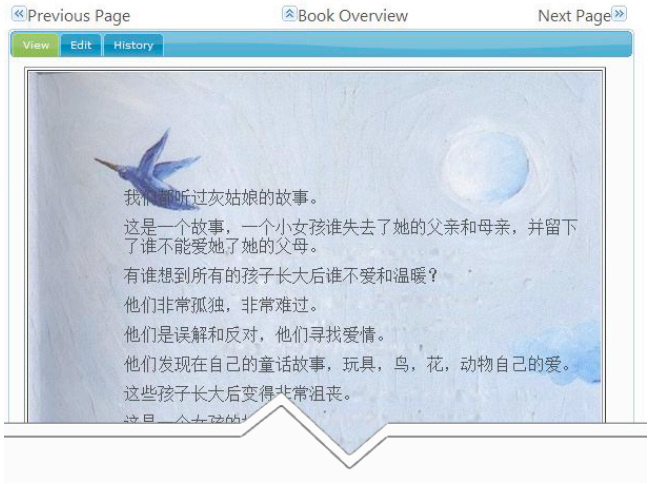

Figure 3. Interface showing the latest revision of Chinese translations (from a children's picture book in English)

4 Technical Realization

We now turn to a fuller discussion of technical realizations for the enrichment channel (a few of which are shown in the figures), underlying automatic MT capabilities, and the interaction of human and machine capabilities.

4.1 Prototype User Interface

Figures 3, 4, and 5 show our prototype of an initial interface design for the monolingual translation protocol illustrated in Figure 2. For illustrative proposes, English and Chinese are used as the source and target languages, respectively. This prototype is structured with the task of book translation in mind, based on the driving need for translation from the International Children's Digital Library project (www.childrenslibrary.org) which has thousands of books in 60 languages that need to be translated [9][27]. However, our software architecture provides a general interchange format that permits translation of other documents such as Wikipedia entries.

When a user enters the translation tool, the UI is in the viewing mode (Figure 3), showing a page in the user's language [28]. For every sentence on the page, the viewing mode shows its most up-to-date translation hypothesis (or corresponding back-translation). In this mode, every translation that needs revision in the current user's language is highlighted. The user can navigate through all available pages with navigation controls, or change into editing mode to edit sentence translations.

In editing mode (Figure 4), sentences are presented in a table for editing convenience. When the user selects a sentence, the UI shows the most up-to-date translation hypothesis with: 1) the last user's revisions (in the other language); 2) detail-on-demand displaying the original source sentence in context; 3) a rich editor where the sentence can be edited and annotated. Once the user finishes editing a sentence, she can send it back to the system, and the system will pass the sentence over to a user speaking the other language. The user can also propose to end the translation process once she is satisfied with the translation quality. When both users agree to end, translation of a sentence is finalized.

The editing mode currently includes the following elements for the enrichment channel, aimed at enhancing redundancy and the communicating of shared context.

Figure 4. Interface showing an ongoing English-to-Chinese translation. When the user selects a sentence from the list (1), its original source is shown (2) with last edit from the partner (3). The user then edits the sentence with the editor (4). The editor also shows the back translation (5).

- **Image annotations.** Images can be associated with a span of text by selecting it in the version of the sentence being worked on, and finding an image with an integrated image search engine. Images are a common way to help bridge the communication gap when linguistic communication is impaired or unavailable [18][23][33].
- **Web link annotations.** Web links can be similarly associated with a span of text by selecting the text and using an integrated search engine to find appropriate links. This feature is especially helpful if translation of the linked page exists. For example, one might annotate the English word *Cinderella* with the link to http://en.wikipedia.org/wiki/Cinderella, which a) identifies the name's translations in a variety of languages, increasing the likelihood that a Chinese speaker with limited knowledge of other languages might recognize it, and b) includes images that help increase shared context.
- **Annotation of correct, incorrect, and uncertain spans.** Unlike conventional translation, our translation protocol provides not only the opportunity to offer the target language user enriched content and broader context, but also the ability for participants to engage in meta-discussion about the translations themselves. That possibility also exists in interactive environments like translated chat or speech-to-speech translation, but the single channel in those settings leads to infinite regress: if a user asks a clarification question, how does he know that the question was translated correctly, and what happens if the answer contains errors, leading to a need for a recursive round of clarification? We avoid this problem in two ways. First, the mechanisms we describe above are designed to add redundancy and increase mutual knowledge, rather than creating language-mediated clarification dialogues. Second, we introduce a limited mechanism for interactive reference to the translation, neither of which requires a recursive step into the translation process.

Figure 5. Interface showing a list of sentences (upper) and history of item 1 (lower)

Participants can indicate spans within the sentence they are revising that have been translated correctly. For example, if the source-language speaker received S_1 = *Everyone loved the story of Cinderella* with the correct source language message S_0 = *Everybody has heard the story of Cinderella*, she might mark "the story of Cinderella" as having been translated correctly before editing S_1 to produce S_2 = *Everyone has heard the story of Cinderella* and send it back through the translation channel. Similarly, a target-language participant receiving *Everyone has heard the business by Cinderella* might mark "the business by" as uncertain, flagging the fact that an error has been detected even if it was not correctable given the available information and making it easier for the source participant to offer relevant clarifying information.

The word level alignments necessary to perform annotation projection can be obtained from our own machine translation engines. Word alignments are also publically available along with translation hypotheses. For example, the Google Translate Research API (opened to university research projects) provides this kind of information [5].

In the UI, there is also a history mode where the iterations of translation can be viewed altogether (Figure 5).

5 Evaluation

5.1 Preliminary Experiment: Wizard of Oz

To begin establishing bounds on the expected quality that our approach might achieve, prior to implementing the prototype we ran a pilot study using a simulated enrichment channel – a "Wizard of Oz" approach – with an actual MT system in the translation channel. In this study, a human "wizard" who knows both languages performed annotation projection in the enrichment channel.

The task in this experiment was to translate several sentences from French to Turkish in a children's book. To ensure that the human participants were effectively monolingual, the experiment involved three languages: one participant knew French and English, the other knew Turkish and English, and the "wizard" knew all three languages. This way, the participants were monolingual with respect to the experimental task (communicating only in French and Turkish, respectively), but the wizard could communicate with both of them, and could also communicate with the experimenters in English as necessary. (We chose French-Turkish translation both to show the potential of our system to work with languages distant from each other, and also based on the availability of a trilingual "wizard".)

During the study, the pair of French-speaking and Turkish-speaking participants communicated according to the iterative protocol only through the system. The experimenter used an MT system and passed a sentence translation (or back-translation) between the two monolingual speakers. Each monolingual speaker only saw the translation (or back-translation) in one language. The kinds of annotations that could be passed through the system included: marking words as correct or ambiguous; adding images; adding Wikipedia links; paraphrasing parts of the sentence; and asking and answering questions with a small set of predefined templates.

The pair worked on five sentences. Of those five, two were translated perfectly, one had a minor error, and two had problems that the pair did not resolve. Since the material in question was a children's picture book, the pictures helped to define a "frame" of possible meaning for the sentences. The monolingual speakers used all the types of annotations offered, relying heavily on annotation of sub-sentential spans as correct or incorrect in order to direct their efforts to parts of the sentences most in need of revision. Paraphrasing was not used explicitly; however, speakers frequently rephrased the sentences to avoid phrases they described as "the machine translation is not good at", which became evident over iterations of the protocol in the form of repeated errors.

This small experiment was both encouraging and challenging. Although the sample size was obviously too small to draw strong conclusions, the process did demonstrate the potential of the protocol to begin with low quality automatic translations and make progress toward high quality outcomes. At the same time, this pilot study made it clear that our initial conception of the protocol can be labor intensive at times, especially for the source-language volunteer; we have addressed this by designing the interface to streamline the interaction, making it as simple as possible to communicate.

5.2 Evaluation of Prototype

After the "Wizard of Oz" experiment, we built the prototype and used it to translate part of a children's book from Russian to Chinese. Chinese and Russian are commonly spoken languages. However, they still make good experimental candidates because they are very different from the perspective of linguistic typology.

Two Russian speakers and four Chinese speakers formed four pairs to use the prototype. (One Russian speaker participated three times, with different content.) The participants were all native speakers of one language and had no knowledge of the other. They were all computer-literate and fluent speakers of English. While most of the participants were computer science students and researchers, none of them work on machine translation directly, and none of them were familiar with the details of this project. They were not linguists or linguistic students.

During the experiment, participants were in the same room but far enough so they could not see each other's screen or hear each other. They were allowed to communicate with the experimenter in English, given the partner would not be able to hear. They were not allowed to (and did not) write anything in English.

Each pair of participants spent an hour together. While they were told to work freely on any sentence (including those that were incompletely translated by previous participants), pairs of participants worked on different pages.

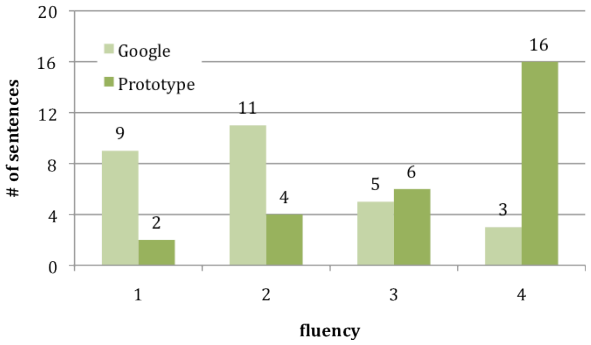

Figure 6. Number of sentences in each fluency category, rated by a professional translator. 1=unintelligible, 4=very intelligible.

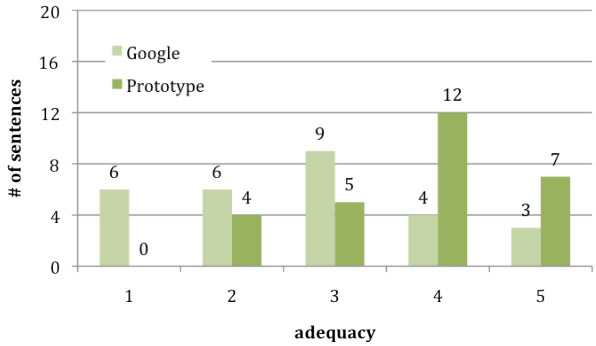

Figure 7. Number of sentences in each adequacy category, rated by a professional translator. 1=not translated, 5=all meaning expressed.

Participants worked on 6 pages (a total of 44 sentences) and finished translating 28 of them. This rate, approximately seven sentences per hour between any given pair of participants, is about five times faster than the "Wizard of Oz" experiment. With a standard rating procedure [14], sixteen of the 28 sentences translated with the prototype were rated as fully fluent and nineteen sentences of the 28 were rated as mostly or fully translated, by a professional translator not connected with the project (see Figures 6 and 7). (There were also incomplete translations with very high quality, but only completed translations are included in the results here.)

The point most worth noting is the shift in adequacy (Figure 7). It is notable that completely inadequate MT outputs (none of the meaning preserved) drop from 6 to 0. This means that the protocol is helping the target language participant understand at least some of the meaning even when the original MT output quality is really low and they have very little to go on. In a coarse-grained way, if the adequacy rating can be categorized so that {1,2}=bad and {4,5}=good, then there is a drop in bad (meaning-wise) from 12 to 4 out of 28, and an increase in good from 7 to 19 of 28; roughly speaking, that represents a factor of 3 in each of the desirable directions.

Note that the improvements in fluency are to be expected given the instructions, as is the heavy non-bell-curve skew toward top fluency. Indeed, anything except a top score in fluency may seem unexpected given the instructions, but natural variation in people's judgments about fluency probably accounts for this.

In addition, we observed promising anecdotal results. For example, two pairs of participants successfully translated their first sentence in five minutes. On every page, more than half of the sentences were successfully translated. Although the prototype has some remaining usability issues, all users correctly understood the iterative protocol. According to participants, the target language speaker's job was to make the best educated guess and the source language speaker's job was to guess if the partner has made the correct guess.

Compared to the "Wizard of Oz" experiment, the prototype offered speed and quality improvements which reconfirmed our design of the protocol. At the same time, it is clear that more work is needed on a number of fronts. For example, during the experiments, our protocol did not work well when sentences were so long and/or difficult that the initial machine translation was very poor. It therefore became very hard for the target language speaker to get a good enough first guess to "bootstrap" the process. There were also cases where the source language speaker could not end the translation process in spite of a correct translation, due to garbled back translations. Morita and Ishida [24] allow users to address these issues by asking their partners to paraphrase the entire sentence; we are seeing promising initial results with a similar but finer grained approach involving automatic detection of translation problems, and elicitation of paraphrases specifically for those problematic regions [11].[2]

6 Discussion

A final issue in the design space for this protocol concerns the nature of the distributed human computation. It is natural to think about the monolingual translation protocol as a synchronous activity that involves two people from beginning to end in translating each sentence. However, since any step can take some time as a person thinks about their work – and of course because two people may not be logged in at the same time, particularly across time zones – the system is designed for asynchronous use, but with good support for synchronous communications when used by two people at the same time who are responding rapidly. To this end, a participant can work on any sentence at any time.

Since the system supports asynchronous participation, it is natural to go a step further, and permit steps in the translation protocol to be distributed among a population of monolingual users. So, for example, French and English speaking populations might step in and out of the participant roles in Figure 1, perhaps contributing only a single (half-)step of an iteration. Distributing participation in this alternative way runs the risk of losing useful context, but, on the other hand, a more fine-grained distribution of human effort would have the advantage of learning from many individuals' perspectives. In addition, units of work could be quite small, and thus it is likely to be easier to recruit participants. Since our interface is designed to preserve context, both synchronous and asynchronous approaches are possible.

Finally, just as (half-)steps in the iterative translation protocol can be distributed as tasks among a population of participants, enrichment operations can also be construed as a "bag of tasks". For example, it is straightforward to create HITs (human intelligence tasks) on Amazon Mechanical Turk for paraphrasing, e.g. presenting "It is a story about a little girl who lost her mother and father" and asking the human worker to replace the underlined phrase with another phrase, without changing the meaning of the sentence [11].

Distributing the translation process will alleviate the problem of long iterations between the same pair of participants. Although the focus of this protocol is to enable translation between uncommon languages, with enough parallelism, the average

[2] These ideas, too, were under development prior to encountering Morita and Ishida's helpful and promising study.

translation speed could also be increased to become comparable with professional translators for common language pairs.

7 Conclusion

Human translators are hard to find, but there are orders of magnitude more monolingual volunteers on the Web than there are translators. To bring high quality translation to scale, we bring together the insights of state-of-the-art MT approaches with the use of distributed human computation, tapping the knowledge of people who speak only one language well.

Guided by principles of redundancy and shared context known to play a significant role in successful linguistic communication, we have designed a protocol that focuses on translation as an iterative, collaborative process between monolingual participants. Our framework makes it possible for human collaborators not only to detect and correct some errors, but also to identify detectable errors that aren't correctable given the current information.

We designed an interface to support this collaborative monolingual translation protocol. Preliminary experiments have shown the potential of this framework.

Going forward, we are looking forward to deploying and testing the interface more widely. In doing so, we will look much more closely at the quality and speed of the generated translations. We also expect to introduce a wider variety of enrichment techniques, and to experiment with machine translations of varying quality.

8 Acknowledgments

This work is supported by the National Science Foundation under grant BCS0941455. The authors would also like to thank Catherine Plaisant for her advice on the French text.

References

[1] Facebook Translation App, 2009.

[2] Category:Available translators in Wikipedia - Wikipedia, the free encyclopedia, 2009.

[3] World Wide Lexicon Toolbar :: Add-ons for Firefox, 2009.

[4] Solve Puzzles for Science | Foldit, 2009.

[5] Documentation – University Research Program for Google Translate - Google Research, 2009.

[6] A.W. Ahmed, A. Badran, A.W. Black, R. Frederking, D. Gates, A. Lavie, L. Levin, K. Lenzo, L.M. Tomokiyo, J. Reichert, T. Schultz, D. Wallace, M. Woszczyna, and J. Zhang, Speechalator: Two-Way Speech-To-Speech Translation In Your Hand, In *Proceedings of the European Conference on Speech Communication and Technology*, pages 369--372, 2003.

[7] L.V. Ahn and L. Dabbish, Designing games with a purpose, In *Commun. ACM*, vol. 51, pages 58-67, 2008.

[8] J. Allen, Post-editing, In *Computers and Translation: A Translators Guide.*, pages 297-317, Amsterdam: John Benjamins., 2003.

[9] B.B. Bederson, Experiencing the International Children's Digital Library, In *interactions*, vol. 15, pages 50-54, 2008.

[10] L. Berlin, A Web That Speaks Your Language, In *The New York Times*, May. 2009.

[11] O. Buzek, P. Resnik, and B.B. Bederson, Error Driven Paraphrase Annotation using Mechanical Turk, In *Creating Speech and Language Data With Amazon's Mechanical Turk, NAACL 2010 Workshop*, 2010.

[12] C. Callison-Burch, Linear B system description for the 2005 NIST MT evaluation exercise, In *Proc. NIST Machine Translation Evaluation Workshop.*, 2005.

[13] H.H. Clark and C.R. Marshall, Definite reference and mutual knowledge, In *Psycholinguistics: Critical Concepts in Psychology*, page 414, 2002.

[14] M. Dabbadie, A. Hartley, M. King, K.J. Miller, W.M. El Hadi, A. Popescu-Belis, F. Reeder, and M. Vanni, A hands-on study of the reliability and coherence of evaluation metrics, In *Workshop at the LREC 2002 Conference*, page 8, Citeseer, 2002.

[15] J. Esteban, J. Lorenzo, A.S. Valderrábanos, and G. Lapalme, TransType2: an innovative computer-assisted translation system, In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 1, Barcelona, Spain: Association for Computational Linguistics, 2004.

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

[17] R. Hwa, P. Resnik, A. Weinberg, and O. Kolak, Evaluating translational correspondence using annotation projection, In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 392-399, Philadelphia, Pennsylvania: Association for Computational Linguistics, 2002.

[18] A. Komlodi, W. Hou, J. Preece, A. Druin, E. Golub, J. Alburo, S. Liao, A. Elkiss, and P. Resnik, Evaluating a cross-cultural children's online book community: Lessons learned for sociability, usability, and cultural exchange, In *Interacting with Computers*, vol. 19, pages 494-511, Jul. 2007.

[19] E. Lagoudaki, Translation memory systems: Enlightening users' perspective, In *Translation Memories Survey 2006*, Imperial College London, 2006.

[20] A. Lavie, F. Metze, R. Cattoni, and E. Costantini, A multi-perspective evaluation of the NESPOLE!: speech-to-speech translation system, In *Proceedings of the ACL-02 workshop on Speech-to-speech translation: algorithms and systems - Volume 7*, pages 121-128, Association for Computational Linguistics, 2002.

[21] M. Maybury, J. Griffith, R. Holland, L. Damianos, Q. Hu, and R. Fish, Virtually Integrated Visionary Intelligence Demonstration (VIVID), In *MITRE technical papers*, 2005.

[22] K. Mieszkowski, I make $1.45 a week and I love it, 2006.

[23] R. Mihalcea and C.W. Leong, Toward communicating simple sentences using pictorial representations, In *Machine Translation*, vol. 22, pages 153-173, 2008.

[24] D. Morita and T. Ishida, Designing Protocols for Collaborative Translation, In *Principles of Practice in Multi-Agent Systems*, pages 17-32, 2009.

[25] R. Prasad, P. Natarajan, D. Stallard, F. Choi, S. Saleem, C. Kao, K. Subramanian, and K. Krstovski, Challenges and Future Directions for Speech-to-Speech Translation, In *Panel on "Problems and Future Directions of Speech Translation Technology"*, 2007.

[26] A.J. Quinn and B.B. Bederson, A Taxonomy of Distributed Human Computation, In *Human-Computer Interaction Lab Tech Report, University of Maryland*, Oct. 2009.

[27] A.J. Quinn, C. Hu, T. Arisaka, A. Rose, and B.B. Bederson, Readability of scanned books in digital libraries, In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 705-714, Florence, Italy: ACM, 2008.

[28] A.J. Quinn, C. Hu, T. Arisaka, A. Rose, and B.B. Bederson, Readability of scanned books in digital libraries, In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 705-714, Florence, Italy: ACM, 2008.

[29] C.E. Shannon, A mathematical theory of communication, In *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 5, pages 3-55, 2001.

[30] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, A Study of Translation Edit Rate with Targeted Human Annotation, In *Proceedings AMTA*, pages 231, 223, 2006.

[31] S. Soderland, C. Lim, B.Q. Mausam, O. Etzioni, and J. Pool, Lemmatic Machine Translation, In *MT Summit XII: proceedings of the twelfth Machine Translation Summit*, pages 128-135., Jul. 2009.

[32] D. Yarowsky, G. Ngai, and R. Wicentowski, Inducing multilingual text analysis tools via robust projection across aligned corpora, In *Proceedings of the first international conference on Human language technology research*, pages 1-8, San Diego: Association for Computational Linguistics, 2001.

[33] X. Zhu, A.B. Goldberg, M. Eldawy, C.R. Dyer, and B. Strock, A text-to-picture synthesis system for augmenting communication, In *Proceedings of the 22nd national conference on Artificial intelligence - Volume 2*, pages 1590-1595, Vancouver, British Columbia, Canada: AAAI Press, 2007.