Correcting Automatic Translations through Collaborations between MT and Monolingual Target-Language Users

Joshua S. Albrecht and **Rebecca Hwa** and **G. Elisabeta Marai**
Department of Computer Science
University of Pittsburgh
{jsa8,hwa,marai}@cs.pitt.edu

Abstract

Machine translation (MT) systems have improved significantly; however, their outputs often contain too many errors to communicate the intended meaning to their users. This paper describes a collaborative approach for mediating between an MT system and users who do not understand the source language and thus cannot easily detect translation mistakes on their own. Through a visualization of multiple linguistic resources, this approach enables the users to correct difficult translation errors and understand translated passages that were otherwise baffling.

1 Introduction

Recent advances in machine translation (MT) have given us some very good translation systems. They can automatically translate between many languages for a variety of texts; and they are widely accessible to the public via the web. The quality of the MT outputs, however, is not reliably high. People who do not understand the source language may be especially baffled by the MT outputs because they have little means to recover from translation mistakes.

The goal of this work is to help *monolingual target-language* users to obtain better translations by enabling them to identify and overcome errors produced by the MT system. We argue for a human-computer collaborative approach because both the users and the MT system have gaps in their abilities that the other could compensate. To facilitate this collaboration, we propose an interface that mediates between the user and the MT system. It manages additional NLP tools for the source language and translation resources so that the user can explore this extra information to gain enough understanding of the source text to correct MT errors. The interactions between the users and the MT system may, in turn, offer researchers insights into the translation process and inspirations for better translation models.

We have conducted an experiment in which we asked non-Chinese speakers to correct the outputs of a Chinese-English MT system for several short passages of different genres. They performed the correction task both with the help of the visualization interface and without. Our experiment addresses the following questions:

- To what extent can the visual interface help the user to understand the source text?
- In what way do factors such as the user's backgrounds, the properties of source text, and the quality of the MT system and other NLP resources impact that understanding?
- What resources or strategies are more helpful to the users? What research directions do these observations suggest in terms of improving the translation models?

Through qualitative and quantitative analysis of the user actions and timing statistics, we have found that users of the interface achieved a more accurate understanding of the source texts and corrected more difficult translation mistakes than those who were given the MT outputs alone. Furthermore, we observed that some users made better use of the interface for certain genres, such as sports news, suggesting that the translation model may be improved by a better integration of document-level contexts.

Proceedings of the 12th Conference of the European Chapter of the ACL, pages 60–68,
Athens, Greece, 30 March – 3 April 2009.

2 Collaborative Translation

The idea of leveraging human-computer collaborations to improve MT is not new; computer-aided translation, for instance, was proposed by Kay (1980). The focus of these efforts has been on improving the performance of professional translators. In contrast, our intended users cannot read the source text.

These users do, however, have the world knowledge and the language model to put together coherent sentences in the target-language. From the MT research perspective, this raises an interesting question: given that they are missing a *translation model*, what would it take to make these users into effective "decoders?" While some translation mistakes are recoverable from a strong language model alone, and some might become readily apparent if one can choose from some possible phrasal translations; the most difficult mistakes may require greater contextual knowledge about the source. Consider the range of translation resources available to an MT decoder–which ones might the users find informative, handicapped as they are for not knowing the source language? Studying the users' interactions with these resources may provide insights into how we might build a better translation model and a better decoder.

In exploring the collaborative approach, the design considerations for facilitating human computer interaction are crucial. We chose to make available relatively few resources to prevent the users from becoming overwhelmed by the options. We also need to determine how to present the information from the resources so that the users can easily interpret them. This is a challenge because the Chinese processing tools and the translation resources are imperfect themselves. The information should be displayed in such a way that conflicting analyses between different resources are highlighted.

3 Prototype Design

We present an overview of our prototype for a collaborative translation interface, named *The Chinese Room*[1]. A screen-shot is shown in Figure 1. It

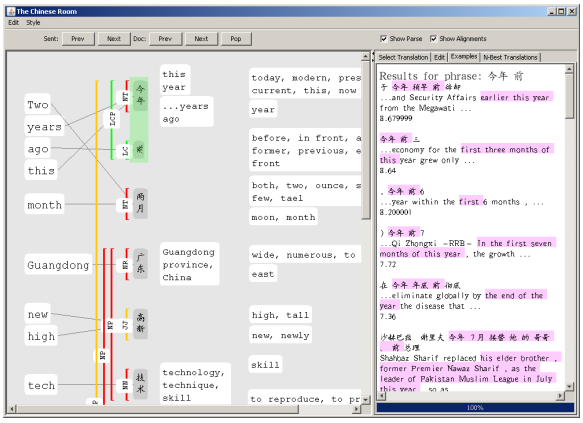

Figure 1: A screen-shot of the visual interface. It consists of two main regions. The left pane is a workspace for users to explore the sentence; the right pane provides multiple tabs that offer additional functionalities.

is a graphical environment that supports five main sources of information and functionalities. The space separates into two regions. On the left pane is a large workspace for the user to explore the source text one sentence at a time. On the right pane are tabbed panels that provide the users with access to a document view of the MT outputs as well as additional functionalities for interpreting the source. In our prototype, the MT output is obtained by querying Google's Translation API[2]. In the interest of exploiting user interactions as a diagnostic tool for improving MT, we chose information sources that are commonly used by modern MT systems.

First, we display the word alignments between MT output and segmented Chinese[3]. Even without knowing the Chinese characters, the users can visually detect potential misalignments and poor word reordering. For instance, the automatic translation shown in Figure 1 begins: *Two years ago this month...* It is fluent but incorrect. The crossed alignments offer users a clue that "two" and "months" should not have been split up. Users can also explore alternative orderings by dragging the English tokens around.

Second, we make available the glosses for words and characters from a bilingual dictionary[4].

[1]The inspiration for the name of our system came from Searle's thought experiment(Searle, 1980). We realize that there are major differences between our system and Searle's description. Importantly, our users get to insert their knowledge rather than purely operate based on instructions. We felt the name was nonetheless evocative in that the user requires additional resources to process the input "squiggles."

[2]http://code.google.com/apis/translate/research

[3]The Chinese segmentation is obtained as a by-product of Google's translation process.

[4]We used the Chinese-English Translation Lexi-

The placement of the word gloss presents a challenge because there are often alternative Chinese segmentations. We place glosses for multi-character words in the column closer to the source. When the user mouses over each definition, the corresponding characters are highlighted, helping the user to notice potential mis-segmentation in the Chinese.

Third, the Chinese sentence is annotated with its parse structure[5]. Constituents are displayed as brackets around the source sentence. They have been color-coded into four major types (noun phrase, verb phrases, prepositional phrases, and other). Users can collapse and expand the brackets to keep the workspace uncluttered as they work through the Chinese sentence. This also indicates to us which fragments held the user's focus.

Fourth, based on previous studies reporting that automatic translations may improve when given decomposed source inputs (Mellebeek et al., 2005), we allow the users to select a substring from the source text for the MT system to translate. We display the N-best alternatives in the *Translation Tab*. The list is kept short; its purpose is less for reranking but more to give the users a sense of the kinds of hypotheses that the MT system is considering.

Fifth, users can select a substring from the source text and search for source sentences from a bilingual corpus and a monolingual corpus that contain phrases similar to the query[6]. The retrieved sentences are displayed in the *Example Tab*. For sentences from the bilingual corpus, human translations for the queried phrase are highlighted. For sentences retrieved from the monolingual corpus, their automatic translations are provided. If the users wished to examine any of the retrieved translation pairs in detail, they can push it onto the sentence workspace.

4 Experimental Methodology

We asked eight non-Chinese speakers to correct the machine translations of four short Chinese passages, with an average length of 11.5 sentences. Two passages are news articles and two are excerpts of a fictional work. Each participant was instructed to correct the translations for one news article and one fictional passage using all the resources made available by *The Chinese Room* and the other two passages without. To keep the experimental conditions as similar as possible, we provided them with a restricted version of the interface (see Figure 2 for a screen-shot) in which all additional functionalities except for the *Document View Tab* are disabled. We assigned each person to alternate between working with the full and the restricted versions of the system; half began without, and the others began with. Thus, every passage received four sets of corrections made collaboratively with the system and four sets of corrections made based solely on the participants' internal language models. All together, there are 184 participant corrected sentences (11.5 sentences $\times$ 4 passages $\times$ 4 participants) for each condition.

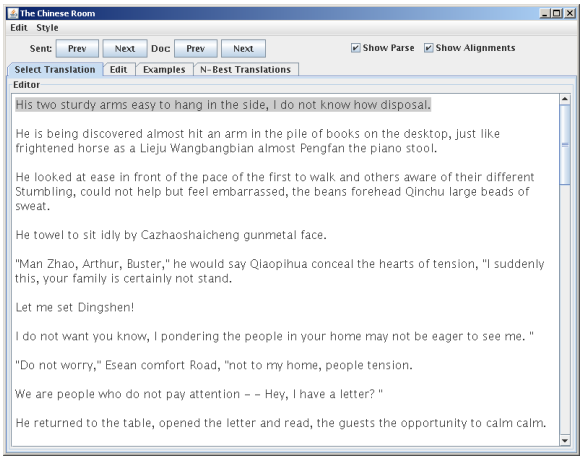

Figure 2: The interface for users who are correcting translations without help; they have access to the document view, but they do not have access to any of the other resources.

The participants were asked to complete each passage in one sitting. Within a passage, they could work on the sentences in any arbitrary order. They could also elect to "pass" any part of a sentence if they found it too difficult to correct. Timing statistics were automatically collected while they made their corrections. We interviewed each participant for qualitative feedbacks after all four passages were corrected.

Next, we asked two bilingual speakers to evaluate all the corrected translations. The outcomes between different groups of users are compared,

con released by the LDC; for a handful of characters that serve as function words, we added the functional definitions using an online dictionary `http://www.mandarintools.com/worddict.html`.

[5]It is automatically generated by the Stanford Parser for Chinese (Klein and Manning, 2003).

[6]We used Lemur (2006) for the information retrieval back-end; the parallel corpus is from the Federal Broadcast Information Service corpus; the monolingual corpus is from the Chinese Gigaword corpus.

and the significance of the difference is determined using the two-sample t-test assuming unequal variances. We require 90% confidence (alpha=0.1) as the cut-off for a difference to be considered statistically significant; when the difference can be established with higher confidence, we report that value. In the following subsections, we describe the conditions of this study in more details.

Participants' Background For this study, we strove to maintain a relatively heterogeneous population; participants were selected to be varied in their exposures to NLP, experiences with foreign languages, as well as their age and gender. A summary of their backgrounds is shown in Table 1.

Prior to the start of the study, the participants received a 20 minute long presentational tutorial about the basic functionalities supported by our system, but they did not have an opportunity to explore the system on their own. This helps us to determine whether our interface is intuitive enough for new users to pick up quickly.

Data The four passages used for this study were chosen to span a range of difficulties and genre types. The easiest of the four is a news article about a new Tamagotchi-like product from Bandai. It was taken from a webpage that offers bilingual news to help Chinese students to learn English. A harder news article is taken from a past NIST Chinese-English MT Evaluation; it is about Michael Jordan's knee injury. For a different genre, we considered two fictional excerpts from the first chapter of *Martin Eden*, a novel by Jack London that has been professionally translated into Chinese[7]. One excerpt featured a short dialog, while the other one was purely descriptive.

Evaluation of Translations Bilingual human judges are presented with the source text as well as the parallel English text for reference. Each judge is then shown a set of candidate translations (the original MT output, an alternative translation by a bilingual speaker, and corrected translations by the participants) in a randomized order. Since the human corrected translations are likely to be fluent, we have instructed the judges to concentrate more on the adequacy of the meaning conveyed. They are asked to rate each sentence on an absolute scale of 1-10 using the guideline in Table 2. To reduce the biases in the rating scales of different judges, we normalized the judges' scores, following standard practices in MT evaluation (Blatz et al., 2003). Post normalization, the correlation coefficient between the judges is 0.64. The final assessment score for each translated sentence is the average of judges' scores, on a scale of 0-1.

[7]We chose an American story so as to not rely on a user's knowledge about Chinese culture. The participants confirmed that they were not familiar with the chosen story.

Table 2: The guideline used by bilingual judges for evaluating the translation quality of the MT outputs and the participants' corrections.

9-10	The meaning of the Chinese sentence is fully conveyed in the translation.
7-8	Most of the meaning is conveyed.
5-6	Misunderstands the sentence in a major way; or has many small mistakes.
3-4	Very little meaning is conveyed.
1-2	The translation makes no sense at all.

5 Results

The results of human evaluations for the user experiment are summarized in Table 3, and the corresponding timing statistics (average minutes spent editing a sentence) is shown in Table 4. We observed that typical MT outputs contain a range of errors. Some are primarily problems in fluency such that the participants who used the restricted interface, which provided no additional resources other than the *Document View Tab*, were still able to improve the MT quality from 0.35 to 0.42. On the other hand, there are also a number of more serious errors that require the participants to gain some level of understanding of the source in order to correct them. The participants who had access to the full collaborative interface were able to improve the quality from 0.35 to 0.53, closing the gap between the MT and the bilingual translations by 36.9%. These differences are all statistically significant (with $>$98% confidence).

The higher quality of corrections does require the participants to put in more time. Overall, the participants took 2.5 times as long when they have the interface than when they do not. This may be partly because the participants have more sources of information to explore and partly because the participants tended to "pass" on fewer sentences. The average Levenshtein edit distance (with words as the atomic unit, and with the score normalized to the interval [0,1]) between the original MT out-

Table 1: A summary of participants' background. ‡User5 recognizes some simple Kanji characters, but does not have enough knowledge to gain any additional information beyond what the MT system and the dictionary already provided.

	User1	User2	User3	User4	User5‡	User6	User7	User8
NLP background	intro	grad	none	none	intro	grad	intro	none
Native English	yes	no	yes	yes	yes	yes	yes	yes
Other Languages	French (beginner)	multiple (fluent)	none	none	Japanese (beginner)	none	none	Greek (beginner)
Gender	M	F	F	M	M	M	F	M
Education	Ugrad	PhD	PhD	Ugrad	Ugrad	PhD	Ugrad	Ugrad

puts and the corrected sentences made by participants using *The Chinese Room* is 0.59; in contrast, the edit distance is shorter, at 0.40, when participants correct MT outputs directly. The timing statistics are informative, but they reflect the interactions of many factors (e.g., the difficulty of the source text, the quality of the machine translation, the background and motivation of the user). Thus, in the next few subsections, we examine how these factors correlate with the quality of the participant corrections.

5.1 Impact of Document Variation

Since the quality of MT varies depending on the difficulty and genre of the source text, we investigate how these factors impact our participants' performances. Columns 3-6 of Table 3 (and Table 4) compare the corrected translations on a per-document basis.

Of the four documents, the baseline MT system performed the best on the product announcement. Because the article is straight-forward, participants found it relatively easy to guess the intended translation. The major obstacle is in detecting and translating Chinese transliteration of Japanese names, which stumped everyone. The quality difference between the two groups of participants on this document was not statistically significant. Relatedly, the difference in the amount of time spent is the smallest for this document; participants using *The Chinese Room* took about 1.5 times longer.

The other news article was much more difficult. The baseline MT made many mistakes, and both groups of participants spent longer on sentences from this article than the others. Although sports news is fairly formulaic, participants who only read MT outputs were baffled, whereas those who had access to additional resources were able to recover from MT errors and produced good quality translations.

Finally, as expected, the two fictional excerpts were the most challenging. Since the participants were not given any information about the story, they also have little context to go on. In both cases, participants who collaborated with *The Chinese Room* made higher quality corrections than those who did not. The difference is statistically significant at 97% confidence for the first excerpt, and 93% confidence for the second. The differences in time spent between the two groups are greater for these passages because the participants who had to make corrections without help tended to give up more often.

5.2 Impact of Participants' Background

We further analyze the results by separating the participants into two groups according to four factors: whether they were familiar with NLP, whether they studied another language, their gender, and their education level.

Exposure to NLP One of our design objectives for *The Chinese Room* is accessibility by a diverse population of end-users, many of whom may not be familiar with human language technologies. To determine how prior knowledge of NLP may impact a user's experience, we analyze the experimental results with respect to the participants' background. In columns 2 and 3 of Table 5, we compare the quality of the corrections made by the two groups. When making corrections on their own, participants who had been exposed to NLP held a significant edge (0.35 vs. 0.47). When both groups of participants used *The Chinese Room*, the difference is reduced (0.51 vs. 0.54) and is not statistically significant. Because all the participants were given the same short tutorial prior to the start of the study, we are optimistic that the interface is intuitive for many users.

None of the other factors distinguished one

Table 3: Averaged human judgments of the translation quality of the four different approaches: automatic MT, corrections by participants without help, corrections by participants using *The Chinese Room*, and translation produced by a bilingual speaker. The second column reports score for all documents; columns 3-6 show the per-document scores.

	Overall	News (product)	News (sports)	Story1	Story2
Machine translation	0.35	0.45	0.30	0.25	0.26
Corrections without The Chinese Room	0.42	0.56	0.35	0.33	0.41
Corrections with The Chinese Room	0.53	0.55	0.62	0.42	0.49
Bilingual translation	0.83	0.83	0.73	0.92	0.88

Table 4: The average amount of time (minutes) participants spent on correcting a sentence.

	Overall	News (product)	News (sports)	Story1	Story2
Corrections without The Chinese Room	2.5	1.9	3.2	2.9	2.3
Corrections with The Chinese Room	6.3	2.9	8.7	6.5	8.5

Table 6: The quality of the corrections produced by four participants using The Chinese Room for the sports news article.

User1	0.57
User2	0.46
User5	0.70
User6	0.73
bilingual translator	0.73

group of participants from the others. The results are summarized in columns 4-9 of Table 5. In each case, the two groups had similar levels of performance, and the differences between their corrections were not statistically significant. This trend holds for both when they were collaborating with the system and when editing on their own.

Prior Knowledge Another factor that may impact the success of the outcome is the user's knowledge about the domain of the source text. An example from our study is the sports news article. Table 6 lists the scores that the four participants who used *The Chinese Room* received for their corrected translations for that passage (averaged over sentences). User5 and User6 were more familiar with the basketball domain; with the help of the system, they produced translations that were comparable to those from the bilingual translator (the differences are not statistically significant).

5.3 Impact of Available Resources

Post-experiment, we asked the participants to describe the strategies they developed for collaborating with the system. Their responses fall into three main categories:

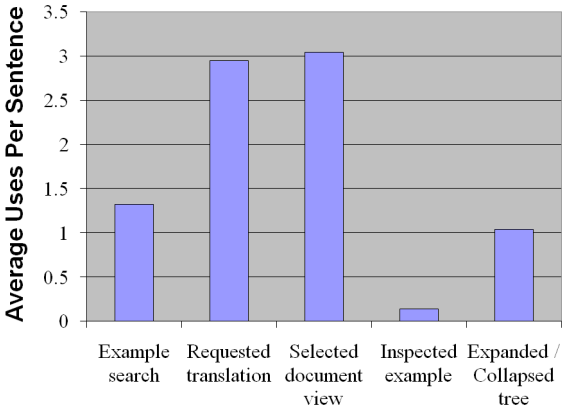

Figure 3: This graph shows the average counts of access per sentence for different resources.

Divide and Conquer Some users found the syntactic trees helpful in identifying phrasal units for N-best re-translations or example searches. For longer sentences, they used the constituent collapse feature to help them reduce clutter and focus on a portion of the sentence.

Example Retrieval Using the search interface, users examined the highlighted query terms to determine whether the MT system made any segmentation errors. Sometimes, they used the examples to arbitrate whether they should trust any of the dictionary glosses or the MT's lexical choices. Typically, though, they did not attempt to inspect the example translations in detail.

Document Coherence and Word Glosses Users often referred to the document view to determine the context for the sentence they are editing. Together with the word glosses and other

Table 5: A comparison of translation quality, grouped by four characteristics of participant backgrounds: their level of exposure to NLP, exposure to another language, their gender, and education level.

	No NLP	NLP	No 2nd Lang.	2nd Lang.	Female	Male	Ugrad	PhD
without The Chinese Room	0.35	0.47	0.41	0.43	0.41	0.43	0.41	0.45
with The Chinese Room	0.51	0.54	0.56	0.51	0.50	0.55	0.52	0.54

resources, the discourse level clues helped to guide users to make better lexical choices than when they made corrections without the full system, relying on sentence coherence alone.

Figure 3 compares the average access counts (per sentence) of different resources (aggregated over all participants and documents). The option of inspect retrieved examples in detail (i.e., bring them up on the sentence workspace) was rarely used. The inspiration for this feature was from work on translation memory (Macklovitch et al., 2000); however, it was not as informative for our participants because they experienced a greater degree of uncertainty than professional translators.

6 Discussion

The results suggest that collaborative translation is a promising approach. Participant experiences were generally positive. Because they felt like they understood the translations better, they did not mind putting in the time to collaborate with the system. Table 7 shows some of the participants' outputs. Although there are some translation errors that cannot be overcome with our current system (e.g., transliterated names), the participants taken as a collective performed surprisingly well. For many mistakes, even when the users cannot correct them, they recognized a problem; and often, one or two managed to intuit the intended meaning with the help of the available resources. As an upper-bound for the effectiveness of the system, we construct a combined "oracle" user out of all 4 users that used the interface for each sentence. The oracle user's average score is 0.70; in contrast, an oracle of users who did not use the system is 0.54 (cf. the MT's overall of 0.35 and the bilingual translator's overall of 0.83). This suggests *The Chinese Room* affords a potential for human-human collaboration as well.

The experiment also made clear some limitations of the current resources. One is domain dependency. Because NLP technologies are typically trained on news corpora, their bias toward the news domain may mislead our users. For example, there is a Chinese character (pronounced *mei3*) that could mean either "beautiful" or "the United States." In one of the passages, the intended translation should have been: *He was responsive to beauty...* but the corresponding MT output was *He was sensitive to the United States...* Although many participants suspected that it was wrong, they were unable to recover from this mistake because the resources (the searchable examples, the part-of-speech tags, and the MT system) did not offer a viable alternative. This suggests that collaborative translation may serve as a useful diagnostic tool to help MT researchers verify ideas about what types of models and data are useful in translation. It may also provide a means of data collection for MT training. To be sure, there are important challenges to be addressed, such as participation incentive and quality assurance, but similar types of collaborative efforts have been shown fruitful in other domains (Cosley et al., 2007). Finally, the statistics of user actions may be useful for translation evaluation. They may be informative features for developing automatic metrics for sentence-level evaluations (Kulesza and Shieber, 2004).

7 Related Work

While there have been many successful computer-aided translation systems both for research and as commercial products (Bowker, 2002; Langlais et al., 2000), collaborative translation has not been as widely explored. Previous efforts such as *DerivTool* (DeNeefe et al., 2005) and *Linear B* (Callison-Burch, 2005) placed stronger emphasis on improving MT. They elicited more in-depth interactions between the users and the MT system's phrase tables. These approaches may be more appropriate for users who are MT researchers themselves. In contrast, our approach focuses on providing intuitive visualization of a variety of information sources for users who may not be MT-savvy. By tracking the types of information they consulted, the portions of translations they selected to modify, and the portions of the source

Table 7: Some examples of translations corrected by the participants and their scores.

	Score	Translation
MT	0.34	He is being discovered almost hit an arm in the pile of books on the desktop, just like frightened horse as a Lieju Wangbangbian almost Pengfan the piano stool.
without The Chinese Room	0.26	Startled, he almost knocked over a pile of book on his desk, just like a frightened horse as a Lieju Wangbangbian almost Pengfan the piano stool.
with The Chinese Room	0.78	He was nervous, and when one of his arms nearly hit a stack of books on the desktop, he startled like a horse, falling back and almost knocking over the piano stool.
Bilingual Translator	0.93	Feeling nervous, he discovered that one of his arms almost hit the pile of books on the table. Like a frightened horse, he stumbled aside, almost turning over a piano stool.
MT	0.50	Bandai Group, a spokeswoman for the U.S. to be SIN-West said: "We want to bring women of all ages that 'the flavor of life'."
without The Chinese Room	0.67	SIN-West, a spokeswoman for the U.S. Bandai Group declared: "We want to bring to women of all ages that 'flavor of life'."
with The Chinese Room	0.68	West, a spokeswoman for the U.S. Toy Manufacturing Group, and soon to be Vice President-said: "We want to bring women of all ages that 'flavor of life'."
Bilingual Translator	0.75	"We wanted to let women of all ages taste the 'flavor of life'," said Bandai's spokeswoman Kasumi Nakanishi.

text they attempted to understand, we may alter the design of our translation model. Our objective is also related to that of cross-language information retrieval (Resnik et al., 2001). This work can be seen as providing the next step in helping users to gain some understanding of the information in the documents once they are retrieved.

By facilitating better collaborations between MT and target-language readers, we can naturally increase human annotated data for exploring alternative MT models. This form of symbiosis is akin to the paradigm proposed by von Ahn and Dabbish (2004). They designed interactive games in which the player generated data could be used to improve image tagging and other classification tasks (von Ahn, 2006). While our interface does not have the entertainment value of a game, its application serves a purpose. Because users are motivated to understand the documents, they may willingly spend time to collaborate and make detailed corrections to MT outputs.

8 Conclusion

We have presented a collaborative approach for mediating between an MT system and monolingual target-language users. The approach encourages users to combine evidences from complementary information sources to infer alternative hypotheses based on their world knowledge. Experimental evidences suggest that the collaborative effort results in better translations than either the original MT or uninformed human edits. Moreover, users who are knowledgeable in the document domain were enabled to correct translations with a quality approaching that of a bilingual speaker. From the participants' feedbacks, we learned that the factors that contributed to their understanding include: document coherence, syntactic constraints, and re-translation at the phrasal level. We believe that the collaborative translation approach can provide insights about the translation process and help to gather training examples for future MT development.

Acknowledgments

This work has been supported by NSF Grants IIS-0710695 and IIS-0745914. We would like to thank Jarrett Billingsley, Ric Crabbe, Joanna Drummund, Nick Farnan, Matt Kaniaris Brian Madden, Karen Thickman, Julia Hockenmaier, Pauline Hwa, and Dorothea Wei for their help with the experiment. We are also grateful to Chris Callison-Burch for discussions about collaborative translations and to Adam Lopez and the anonymous reviewers for their comments and suggestions on this paper.

References

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2003. Confidence estimation for machine translation. Technical Report Natural Language Engineering Workshop Final Report, Johns Hopkins University.

Lynne Bowker. 2002. *Computer-Aided Translation Technology*. University of Ottawa Press, Ottawa, Canada.

Chris Callison-Burch. 2005. Linear B System description for the 2005 NIST MT Evaluation. In *The Proceedings of Machine Translation Evaluation Workshop*.

Dan Cosley, Dan Frankowski, Loren Terveen, and John Riedl. 2007. Suggestbot: using intelligent task routing to help people find work in wikipedia. In *IUI '07: Proceedings of the 12th international conference on Intelligent user interfaces*, pages 32–41.

Steve DeNeefe, Kevin Knight, and Hayward H. Chan. 2005. Interactively exploring a machine translation model. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 97–100, Ann Arbor, Michigan, June.

Martin Kay. 1980. The proper place of men and machines in language translation. Technical Report CSL-80-11, Xerox. Later reprinted in *Machine Translation*, vol. 12 no.(1-2), 1997.

Dan Klein and Christopher D. Manning. 2003. Fast exact inference with a factored model for natural language parsing. *Advances in Neural Information Processing Systems*, 15.

Alex Kulesza and Stuart M. Shieber. 2004. A learning approach to improving sentence-level MT evaluation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, Baltimore, MD, October.

Philippe Langlais, George Foster, and Guy Lapalme. 2000. Transtype: a computer-aided translation typing system. In *Workshop on Embedded Machine Translation Systems*, pages 46–51, May.

Lemur. 2006. Lemur toolkit for language modeling and information retrieval. The Lemur Project is a collaborative project between CMU and UMASS.

Elliott Macklovitch, Michel Simard, and Philippe Langlais. 2000. Transsearch: A free translation memory on the world wide web. In *Proceedings of the Second International Conference on Language Resources & Evaluation (LREC)*.

Bart Mellebeek, Anna Khasin, Josef van Genabith, and Andy Way. 2005. Transbooster: Boosting the performance of wide-coverage machine translation systems. In *Proceedings of the 10th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 189–197.

Philip S. Resnik, Douglas W. Oard, and Gina-Anne Levow. 2001. Improved cross-language retrieval using backoff translation. In *Human Language Technology Conference (HLT-2001)*, San Diego, CA, March.

John R. Searle. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3:417–457.

Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326, New York, NY, USA. ACM.

Luis von Ahn. 2006. Games with a purpose. *Computer*, 39(6):92–94.