Improving Machine Translation by Showing Two Outputs

Bin Xu[1] **Ge Gao**[2] **Susan R. Fussell**[1,2] **Dan Cosley**[1]
[1]Department of information Science
Cornell University
Ithaca NY 14850 USA
[2]Department of Communication
Cornell University
Ithaca NY 14850 USA
{bx55, gg365, sfussell, drc44}@cornell.edu

ABSTRACT
We propose to improve real-time communication between people who do not share a common language by foregrounding potential problems in machine translation. We developed a prototype chat tool that displays two parallel translations of each chat turn, with the thought that comparing the translations might both highlight problems and provide resources for resolving them. We conducted a user study to investigate how people use and like such an interface compared to a standard one-translation interface. On balance, users preferred two translations to one, using them to both notice differences and infer meaning from uncertain translations, with no increase in workload. This suggests that this interface may help improve cross-lingual communication in practical applications and lays the groundwork for a larger design space around systems that highlight possible errors to support communication.

Author Keywords
Machine translation; multiple translations

ACM Classification Keywords
H.5.3. Information interfaces and presentation (e.g., HCI): Group and Organization Interfaces.

INTRODUCTION AND BACKGROUND

Machine translation (MT) systems have been used to support cross-lingual communication for decades, with the goal that one day, high-quality machine translation services will support transparent conversation between people who share no language in common. However, translation is not there yet, especially in unrestricted domains such as the kind of informal communication that is a cornerstone of building trust [1]. MT services still frequently make errors and lack both the specific and the cultural context of the conversation, causing asymmetries and inconsistencies [8] and reducing their utility for cross-lingual communications.

Still, MT systems may help bridge language barriers when human translators are unavailable or expensive, and there are many real-world cases, from tourism to teamwork, where there is little or no linguistic common ground. Thus, much work has gone into improving the algorithms behind these systems. However, errors persist, and interfaces that simply present translated text hide the fact that there are alternative word translations, alignments, and so on.

Showing these alternatives may have value. Studies on back-translation systems show that retranslating a speaker's translated messages back to the speaker's language improves awareness of how messages are processed by MT, thus improving communication quality [7]. Crowdsourcing systems like Monotrans [5] and DuoLingo use human judgment to find and correct flaws in MT by iterating through translations. Google's Translate interface allows users to view alternate translations for individual words and phrases and also shows alignments between elements of the original and the translated text.

THE IDEA: SHOWING MULTIPLE TRANSLATIONS

These ideas suggest that rather than hiding their flaws, MT systems might foreground them in ways that make the problems easier to see, exposing seams in the technology [5] and providing resources to better use the translations. In this paper, inspired by ensemble learning approaches and by the saying "two heads are better than one", we explore the idea of showing multiple parallel translations generated by different engines. Our hope is that this will allow CMC (computer mediated communication) systems that use MT services can implicitly provide information about confidence and alternatives that people can interpret based on redundancies and differences between the translations.

We first explored this idea by translating both English and Chinese conversational turns to the other language using the Google, Bing, and Youdao engines and showing them to both Chinese and English-speaking members of the research lab. In general, we found that when the translations are redundant, either in part or completely, those parts of the translation tended to be both syntactically and semantically correct. Thus, we expect that reading

CHI 2014, April 26 - May 01 2014, Toronto, ON, Canada

ACM 978-1-4503-2473-1/14/04 $15.00.
http://dx.doi.org/10.1145/2556288.2557171

redundant text will increase people's confidence that they understand their partner's intent.

When the translations are not redundant, the situation is more complicated. We expect people will often be able to infer their partner's intent by using redundant aspects to align the translations and their knowledge of the conversational context to choose elements from both translations that make the most sense.

At other times, especially as translation quality declines, they won't be able to make sense of the two translations at all. In this case, we expect there to be some benefit because the fact that something amiss with the translations is clearer than if there is only one translation, and people will realize this and initiate clarifications more quickly when they are needed—but there will be cost in processing the extra text.

Finally, there will be times when one translation is high quality and one is low quality, and in these cases showing two translations is pure cost with no benefit.

We expect that the relative frequency of these different cases will depend on the particular translation engines, language pairs, topics of conversation, and phrasing and word choices, and this will affect the usefulness of the system. On balance, we expect that with two translations rather than with one, people will be more confident in both the partner's intent and the progress of the conversation, and that this will be worth the effort of reading an additional translation. This, in turn, should lead to less clarification work, quicker clarification when it is needed, and smoother conversations overall.

Experimental prototype

We developed a prototype to present our idea in the context of real-time conversation supported by MT. Our scenario was to support chatting between partners who do not both have enough fluency in a common language to have an effective conversation in it. The system has two main parts, the messenger window and the MT backend. We chose to present only two translations because in pilots people reported three translations caused much more workload than two, and chose Google and Bing because they are competitive in quality while using distinct algorithms [9].

When one person sends a message, it first goes to our own server, which forwards the request to Google and Bing. Once the translations are received it sends them to the partner, whose client displays them, always in the same order so that people would get used to the characteristics of each engine. In pilot testing there was no additional noticeable delay introduced by generating two translations instead of one.

USER STUDY

We conducted a within subject user study to investigate how people use and react to showing two translations. Each participant used both the two-translation interface described above and a one-translation interface that was identical except that it presented only one translation. For the one-translation version we used Google, because it is slightly better than other engines in Chinese-English translation [9].

Tasks and data collection

We designed a chatting task to represent a typical informal conversation scenario that might occur when meeting new people either casually or in new work teams. Because it is hard to find Chinese students who can't speak at least some English at a U.S. university, in this study we focused on only the English speakers and used two Chinese confederates from Mainland China to be their partners. They were trained to conduct conversations in a natural, consistent way without steering the conversation.

Participants were instructed to discuss at least three things about themselves and learn at least three things about their partners; we suggested six possible topics but allowed them to use others as well. Each participant spoke to each confederate in two separate sessions; in one they used the single translation interface and in the other they used the two-translation version. We counterbalanced the order of confederates and interfaces.

During the task participants followed a think-aloud protocol. We explained that the goal of our study was to evaluate the system and that we would record their voice and their chatting on the computer simultaneously, and asked them to say aloud everything that went though their mind. Confederates wore headphones so as not to hear participants' voices. We asked partners to chat for 15 minutes in each session; the average time for the single- and two-translation cases was 16.5 minutes (SD=3.1) and 15 minutes (SD=5.2) respectively. After each session, there was a survey asking about their experience during the task, including workload (based on the NASA TLX scale [4]).

After completing both sessions, participants filled out a short post-survey asking their preference between the two interfaces and strategies they used when making sense of the two translations, along with a short (3-5 minute) interview about how they felt about two interfaces, preferences between them, strategies for using them, and suggestions for improvement.

We used an online research recruitment website to recruit 8 participants (5 female, 3 male; ages 18-23 with a median of 20), who were all U.S. citizens at a large northeastern U.S. university. They were native English speakers unable to speak Mandarin Chinese and in the pre-survey rated themselves as frequent users of IM chatting tools, but not of MT tools. Quotes are labeled by participant number.

RESULTS

Figure 1(a) shows the two-translation chat window. Figure 1 also illustrates several ways translation pairs varied for participants: being similar like the first pair in (a), being partly different like "blue Berry" vs. "blueberry" in

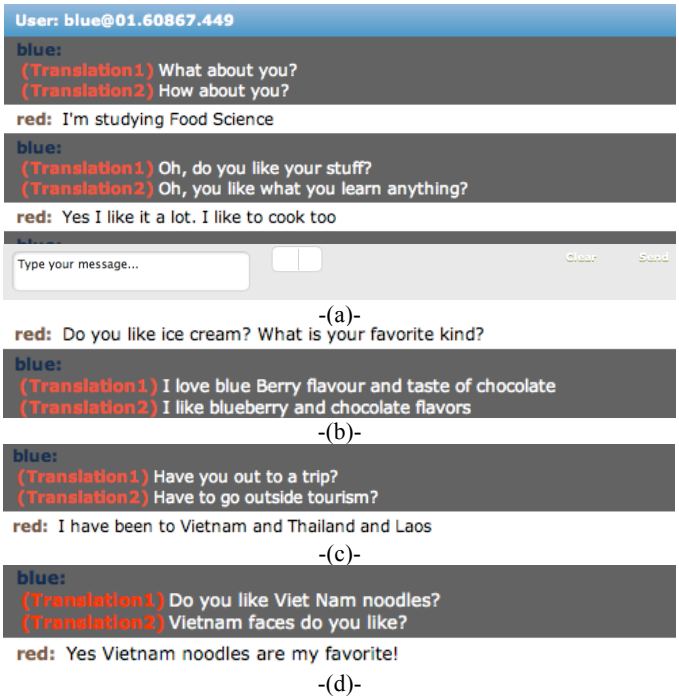

Figure 1. (a) Snapshot of the chat window in the two-translation case, with three additional samples (b, c and d).

(b), being different and one has information being absent in the other one, like "learn anything" in (a)'s second pair, being different but semantically related like "trip" and "tourism" in (c), and finally being different and one is higher quality like the pair in (d). We computed the Levenshtein string edit distance between translations, where small distances usually correspond to similar translations and larger distances tend to be more distinct. The average difference was 36% of the length of the translation, which means that engines often generated quite different translations even on these short, informal texts.

People overall preferred two translations

Interviews showed that 7 of the 8 participants preferred the two-translation interface: *"I definitely like the double translation better than the single translation. In cases where the message is a little bit confusing or vague, the fact that you have a second option to look at is pretty nice."* (P5) and *"I prefer the double …if one translator didn't get it maybe the other one could have provided more clarity."* (P4) This is corroborated by participants' ratings of the usefulness of having a second translation when one was confusing as 6.1 (SD=0.99) on a 7-point Likert scale (1 is least helpful), demonstrating fairly strong liking for it. All participants reported using both translations at least some of the time on the survey.

Differences invite questions, and sometimes choices

We expected that by examining both translations people would become more aware of both specific problems with particular translations and of the potential for errors. This tended to happen when the translations were different: participants in the think-aloud process would often read the two translations, but only when they were different. Often they then just composed a reply, but some differences led to further reflection. For example, when P1 saw Figure 1(b), she said, *"It is kind of interesting the 'blueberry' in translation 1 was separated and 'Berry' was capitalized, where the translation 2 just had 'blueberry'. And the chocolate was 'taste of chocolate' and 'chocolate flavors'."*

This often took the form of thinking about the quality or accuracy of the two translations. For example, when P5 saw "You now what subject?" and "What is your current discipline?" he said, *"The second translation is definitely more accurate, probably is towards what the other person wants to say …you know 'what subject' really does not make any sense, whatsoever, so 'what is your current discipline' is definitely a valid question."*

This idea of comparing the translations, and looking for differences, was a persistent theme in how people thought about the interface: *"…while sometimes they were fairly similar …sometimes they were different enough that I could look at the two and compare them."* (P7)

Differences also help people infer meaning

The snippet from P5 above shows that when one translation is clearly better, people tended to rely on it. However, people sometimes also use parts of both translations, with the differences helping people guess about their partner's intent: *"…sometimes the organization of words was different and sometimes that would help me sort of come to a meaning, understanding of what they're trying to communicate."* (P6) For example, when P1 saw the second pair in Figure 1(a), she said, *"…wording stuff is different. And one says 'do you like your stuff'…and the other is 'oh you like what you learn anything'. I mean, between the two it was easy to figure out what is being asked (about study)."*

This was seen as useful even when the differences were small: *"I think the two translations was a lot easier to use, even though sometimes they didn't vary by much, but sometimes that little variation helped a lot."* (P1)

Even when both translations were flawed, people could often still use them if the flaws were complementary: *"…sometimes they'd both be kind of incorrect but in different ways so I think it's good to have them both kind of work together so you can get what it is they're trying to say."* (P8) In these cases, people might not try to infer the exact, best translation, but to have a good enough idea of the partner's intent to continue the conversation. Often there was enough commonality or redundancy to make this guess. For example, when P1 saw the pair in Figure 1(c), she said *"I got a question, 'have you out to a trip', and the second is 'have to go outside tourism'. Obviously it is about trip, about touring"* to help her infer topics. However, one participant found having two flawed translations to be more confusing than one: *"I think with two, if they were both confusing then it was hard for me to judge which one was*

more correct, but if it were just one I could try to figure out, I think, what they meant." (P3)

Similar translations are not so helpful

We had thought that participants would also benefit from redundancy in translations that would give them confidence that the redundant parts were correct. However, we did not see much evidence of this. Instead, when the translations were the same it was not helpful: *"…it was definitely better when they were different, when they would return different results, to use them both. …if they were the same then that would make me feel more confident, but most of the time the results when they were the same, even when they were still fairly correct they still both had like the same errors."* (P7).

Two translations were not harder than one

We measured users' response time to partners and found no significant difference between the one- and two-translation conditions on average (14.7 vs. 15.2 seconds respectively). Likewise, on the self-reported NASA TLX workload questions, there was no significant difference between the conditions, and participants reported no workload concerns during post-experiment interviews.

DISCUSSION AND CONCLUSION

The lack of difference in workload, participants' preference for two translations over one, and their feedback that differences between translations helped them infer partner's meanings all suggest that showing two translations has promise, at least in this population, setting, and task. Our guess is that the workload was perceived as no higher because extra effort spent on reading two translations was made up for by the gains in understanding they provided.

More study will be needed to identify the tasks, language pairs, and other characteristics of conversational contexts that make showing two translations more and less valuable. More generally, as second language fluency improves, the value of two translations (and MT in general) eventually will fall below just conversing in a common language. Still, there may be times or topics where a person would rather express themselves in their native language and use MT tools to supplement their own proficiency.

In those cases strategies such as this that help people get the most out of MT will still matter. Multiple translations of words (as Google Translate provides) or whole turns, back-translations, keyword highlighting [3] and other visual clues [2] all may help people detect and recover from problems by foregrounding flaws in ways that also provide resources for recovery. In the case of showing two translations, the implementation is not complex and the value is good, suggesting that this is a practical idea for real systems.

We also see this as an interesting case of the general idea of foregrounding system flaws. As technology improves, these strategies may not matter for MT: *"…it would be really nice if the computer could figure out eventually and only give you one message because it'll be easier to talk to the person because it'll be faster: you wouldn't have to read two messages, it would be one, and then you could message quicker and it'll be more fluid."* (P5)

But for now, *"it's still really nice to have two messages, like I speak with other people in different like in Russian and stuff and like often it's not very accurate and so having two translations helps things, so that's definitely cool."* (P5) More generally, we hope this case encourages other designers of tools where the technology plays an active, but imperfect, role in shaping communication [5] to think about ways the interface can help users make the most of it.

ACKNOWLEDGMENTS

This research was funded by NSF grant #1318899. We also thank Xi Yang, Cen Guo, Bridget Rudgers, Ivy Huang, Leslie Setlock, and Tina Yuan for their contributions.

REFERENCES

1. Chevrier, S. Cross-cultural management in multinational project groups. *J. World Business*, *38*, (2003), 141–149.
2. Finch, A., Song, W., Tanaka-Ishii, K., & Sumita, E. Picotrans. Using pictures as input for machine translation on mobile devices. In *Proc. IJCAI 2011*. AAAI Press (2011), 2614–2619.
3. Gao, G., Wang, H.-C., Cosley, D., & Fussell, S. R. Same translation but different experience: the effects of highlighting on machine-translated conversations. In *Proc. CHI 2013,* ACM Press (2013), 449–458.
4. Hart, S. G., & Staveland, L. E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Human mental workload*, 1, (1988), 139–183.
5. Hu, C., Bederson, B. B., Resnik, P., & Kronrod, Y. Monotrans2: A new human computation system to support monolingual translation. In *Proc. CHI 2011*, ACM Press (2011), 1133–1136.
6. Sengers, P., & Gaver, B. Staying open to interpretation: engaging multiple meanings in design and evaluation. In *Proc. DIS* 2006, ACM Press (2006), 99–108.
7. Shigenobu, T. Evaluation and usability of back translation for intercultural communication. *Usability and Internationalization. Global and Local User Interfaces*. Springer (2007), 259–265.
8. Yamashita, N., & Ishida, T. Effects of machine translation on collaborative work. In *Proc. CSCW 2006,* ACM Press (2006), 515–524.
9. Zhang, R. P., Pan, Y., & Yang, Y. (2012). A comparative case study of Google and Bing translation. In *Proc. ICERI 2012*, 3669–3673.