MonoTrans2: A New Human Computation System to Support Monolingual Translation

Chang Hu\textsuperscript{1,2,3}, Benjamin B. Bederson\textsuperscript{1,2,3}, Philip Resnik\textsuperscript{1,3,4} and Yakov Kronrod\textsuperscript{3,4}

\textsuperscript{1}Computer Science Department, \textsuperscript{2}Human-Computer Interaction Lab, \textsuperscript{3}Institute for Advanced Computer Studies, \textsuperscript{4}Department of Linguistics, University of Maryland

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

ABSTRACT
In this paper, we present MonoTrans2, a new user interface to support monolingual translation; that is, translation by people who speak only the source or target language, but not both. Compared to previous systems, MonoTrans2 supports multiple edits in parallel, and shorter tasks with less translation context. In an experiment translating children's books, we show that MonoTrans2 is able to substantially close the gap between machine translation and human bilingual translations. The percentage of sentences rated 5 out of 5 for fluency and adequacy by both bilingual evaluators in our study increased from 10\% for Google Translate output to 68\% for MonoTrans2.

Author Keywords

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms
Design, Human Factors, Experimentation

INTRODUCTION
Monolingual translation, or translation by people who speak only the source or the target language, has been proposed as a way to solve the problem of translating between rare languages, or to achieve more cost effective, high quality translation at a large scale [2][3]. At the core of monolingual translation are protocols in which the human participants (monolingual source or target language speakers) work together to make sense of machine translations. Since monolingual translation does not depend on bilingual humans, it enables translation between uncommon language pairs where a bilingual translator is hard to find. In addition, monolingual translation can be supported by a larger population, and thus it is likely to result in higher throughput. In scenarios where bilingual participants are still required to guarantee quality, it can transform their role into a combination of translation checking and translation creation, rather than just the latter.

Our previous monolingual translation system, MonoTrans, showed promising results [2]. However, we encountered a number of limits to its scalability. First, MonoTrans requires users to participate in a strictly alternating sequence, thus requiring the presence of speakers on both sides for a sustained period for a successful and timely translation. Second, to successfully complete translations, users need to understand the different roles of source and target language speakers, which can be confusing.

These limitations render the system impractical, because participant populations are usually imbalanced. There can be many source language speakers and very few target language speakers, or vice versa. Those users may have vast differences in their schedules and motivations; e.g., some may only intend to spend too short a time to understand the translation UI.

Drawing upon this experience, we present MonoTrans2, a new monolingual translation system, with two goals:

1) Participation in parallel: Enable every participant to take part in the translation process at any time.

2) Minimal context: Decrease the amount of contextual knowledge required to participate.

We performed an experiment in which MonoTrans2 was deployed to translate children's books in the International Children’s Digital Library (www.childrenslibrary.org, ICDL). The results show a dramatic improvement in the production of high quality translations (as determined by bilingual evaluators), without relying on any human bilingual expertise.

RELATED WORK
Monolingual translation is a fairly new idea. Previous systems include MonoTran [2] and the Language Grid system [3]. These are examples of Human Computation, systems that connect people and computers to solve problems not easily addressable by either one alone [4]. In these systems, a machine translation channel connects one or more source language speakers with target language speakers. On each side, the users perform actions such as edit, accept, reject, and/or attach extra information. The source and target speakers try to make sense of the (usually
The Language Grid experimental system [3] is based on a two-phase synchronous protocol involving two users. In the first phase, the source language speaker repeatedly rephrases the sentence being input to the MT system until the target language speaker indicates that the machine translation can be understood. Then the protocol enters the second phase, in which the target language speaker proposes one rephrasing of the machine translation at a time, until the source language speaker decides that the translation is sufficient by comparing the back-translation with the original. The Language Grid system assumes that users are online and that the same pair (or pairs) of users are working on the same sentence extensively until it has been translated.

MonoTrans [2] was based on a different iterative protocol. Similarly, the source sentence is first automatically translated into the target language and given to the target language speaker. Then, at each step, the source and the target language speaker take turns to edit or attach extra information to the translation (or its back-translation). MonoTrans has a clearly defined back-and-forth process where each side's task must be completed before the next step of work can be done.

**Table 1. Tasks in MonoTrans2**

<table>
<thead>
<tr>
<th>Source Side</th>
<th>Target Side</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vote on a candidate</td>
<td>Vote on a candidate</td>
</tr>
<tr>
<td>Give explanation to marked &quot;error&quot; phrases</td>
<td>Mark a phrase in the candidate as &quot;error&quot;</td>
</tr>
<tr>
<td>Rephrase the source sentence</td>
<td>Suggest a new translation candidate</td>
</tr>
</tbody>
</table>

We designed MonoTrans2 (Figure 1) to support participation in parallel as well as tasks with minimal context. Like all the previous systems, MonoTrans2 still connects two (groups of) users through a machine translation channel. However, it allows translation or back-translation to be edited by multiple users on the same side simultaneously. The edited (back-)translations are then sent to users on the other side in parallel. Compared with translation in MonoTrans, the biggest difference is that now there are multiple parallel threads of translations/back-translations for any one sentence. This enables multiple users to participate at the same time without waiting.

To deal with multiple parallel translation threads, candidates and voting are introduced. The simultaneous edited (back-)translations are presented in MonoTrans2 as translation candidates. Target language speakers propose translation candidates (Figure 2). These candidates are back-translated and presented to the source language speakers in parallel with each other. Source language speakers, in turn, can propose a rephrasing of the original sentence at any time, and the rephrased source sentences become candidates, which are then translated and presented to the target language speakers. To identify the best candidate, MonoTrans2 allows users to vote for or against any candidate. Voting not only helps to identify the best candidate in a cumulative manner, but also provides casual users with a very short and intuitive task.

There are six tasks in MonoTrans2, three tasks for each side (see Table 1.) During the design of these tasks, special attention was paid to minimizing users’ need for context. For example, we redesigned the mechanism that allows users to attach clarifying information (such as an image or URL) to parts of a sentence. In an initial design of MonoTrans2, target language speakers asked questions about the parts they felt needed clarification, and source language speakers attached clarifying information as answers. This design created a tension for casual users who use MonoTrans only once on the target side, because they either send questions for which they will never see an answer, or receive answers for questions they did not ask. To avoid this problem, this functionality was refactored so that target language speakers only mark parts of sentences that need clarification (without asking a specific question), which is a very simple and intuitive task, and source language speakers see which parts of the source sentence
were confusing and choose the clarification type and content.

**EXPERIMENT**

We evaluated MonoTrans2 by translating content from children’s books (intended reader’s age 6-9) from the ICDL. We had volunteers work on translating four Spanish books into German, and one German book into Spanish. All translations started from the language in which the book was originally published. Machine translation was done using the Google Translate Research API.

Participants were recruited from a database of ICDL volunteer translators, soliciting people who spoke German or Spanish (but not both). Sixty (60) fluent Spanish speakers and 22 fluent German speakers participated. In four days, participants edited/voted on 162 sentences. We estimate 20 words per typical sentence, so on average, roughly 800 words were translated per day, which is close to half the speed of professional translators.\(^1\)

Two fluently bilingual evaluators unfamiliar with the project were recruited to assess translation quality for fully automatic output of Google Translate (as a baseline) and for the output of MonoTrans2 (using Google Translate as the translation channel). They were paid for their work. Systems were identified only as Systems A and B, and evaluators were not told which outputs came from which system. For each output (target sentence) paired with its corresponding source sentence, the evaluator’s task was to rate the target sentence’s fluency and adequacy on a 5-point scale, where fluency of 5 indicates complete fluency and adequacy of 5 indicates complete preservation of meaning [1]. Bilingual assessments were obtained for all five books.

In addition, two native Spanish speakers, both unfamiliar with German, rated quality for translations of the German book into Spanish. They also were independent of the project and paid for their time. They were given ground-truth Spanish translations for reference. The evaluation results are shown in Table 2.

Figures 3 and 4 summarize the results. Unsurprisingly, MonoTrans2 produced large gains in fluency compared with raw machine translation output.\(^2\) Crucially, however, note the shift in adequacy in Figure 4: using MonoTrans2, the peak of the adequacy distribution has been shifted from 3 to 5. We ran four two-tailed paired t-tests, between scores of MonoTrans2 and scores of Google Translate for both fluency and adequacy for each bilingual evaluator (Table 4). We also ran chi-square tests for the scores (both for individual evaluators, and for the evaluators in aggregate). The p values are well under .001 (Table 5). These results show that MonoTrans2, using only monolinguals, significantly improved translation fluency and adequacy over Google Translate.

<table>
<thead>
<tr>
<th>Google</th>
<th>MonoTrans2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sentences with fluency = 5</strong></td>
<td>21 (13%)</td>
</tr>
<tr>
<td><strong>Sentences with adequacy = 5</strong></td>
<td>17 (10%)</td>
</tr>
<tr>
<td><strong>Sentences where BOTH = 5</strong></td>
<td>17 (10%)</td>
</tr>
</tbody>
</table>

**Table 3. Number of sentences with maximum possible fluency and adequacy. Number of sentences, N = 162.**

Table 3 conveys our experiment’s bottom-line results more strikingly. On the very conservative criterion that a translation output is considered high quality only if *both* bilingual evaluators rated it a 5 for *both* fluency and adequacy, Google Translate produced high quality output for 10% of the sentences, while the MonoTrans2 process improved this to 68%.

We also elicited subjective reactions from our monolingual translation participants. In the fifteen written comments, the greatest concern was that we might simply publish the translations created through our process without bilingual review; as ICDL volunteers, these participants were deeply committed to attaining 100% bilingual-translator quality. In followup correspondence, their concerns were alleviated once they understood that in the real-world ICDL use case,\(^1\)

\(^1\) Generally, professional translators’ speed is 2000 words per day. We hired a translation firm to translate the same books. They took four days.

\(^2\) Improvements in fluency are to be expected given that the last step of output is always produced by a human, as is the heavy non-bell-curve skew toward top fluency.
the results suggest MonoTrans2 could potentially convert 68% of bilingual translators’ time to validation rather than full translation, but the role of those bilinguals would remain necessary.

### Evaluation

<table>
<thead>
<tr>
<th>Evaluator</th>
<th>Fluency</th>
<th>Adequacy</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>63e-16</td>
<td>3e-19</td>
</tr>
<tr>
<td>B2</td>
<td>5e-27</td>
<td>6e-27</td>
</tr>
</tbody>
</table>

Table 4. T-test results (p values)

<table>
<thead>
<tr>
<th>Evaluator</th>
<th>Fluency</th>
<th>Adequacy</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>89.4</td>
<td>85.1</td>
</tr>
<tr>
<td>B2</td>
<td>115</td>
<td>115</td>
</tr>
<tr>
<td>Both</td>
<td>192</td>
<td>192</td>
</tr>
</tbody>
</table>

Table 5. Chi-square test results (F values, DOF=4)

**DISCUSSION**

MonoTrans2 users appeared to finish more work in the same amount of time than users of the original MonoTrans. However, there were a number of differences in these studies, so this should be taken only as a rough comparison. While one might expect finer-grained comparisons between the two systems, the value of such a comparison would be unclear, since MonoTrans2 is not an incremental revision, but rather a new design inspired by, and drawing on the old design. So much has been changed that a direct comparison with MonoTrans would not provide the ability to isolate particular dimensions of the design, and the processes are so different that they do not offer any obviously comparable user-level measures. In this paper, the key, bottom-line result is the comparison between fully automatic translation and the new process: a dramatic improvement in the production of high quality translations, without requiring any human bilingual expertise.

**CONCLUSION**

This paper presented MonoTrans2, a new monolingual translation system that supports participants working in parallel with minimal context. Our experiment confirmed that tasks in a translation process can be designed so each user performs tasks independently of the others, rather than working synchronously with another monolingual user as in every previous monolingual translation system. Breaking the monolinguals’ tasks down and shortening them ensures that there are always tasks available for all users, making the approach more scalable and more attractive to participants regardless of their intended time commitment. In comparison with state of the art automatic machine translation, MonoTrans2 obtains dramatic improvement in the production of high quality translations, without relying on any human bilingual expertise. The approach shows significant promise for use cases like ICDL, where bilingual costs and effort must be kept to a minimum. In addition to fully integrating bilingual post-processing in the ICDL scenario, we also plan to evaluate the approach in other scenarios where bilingual validation simply cannot be expected.

**ACKNOWLEDGMENTS**

This research is supported by NSF contract #BCS0941455 and by a Google Research Award.

**REFERENCES**


---

3 In an unpublished study of MonoTrans, 22 participants translated 12 sentences in 4 days.