% !TEX root = paper.tex

% !TEX encoding = UTF-8 Unicode

\input{figure1}

\section{Results}

\label{sec:results}

\subsection{Rating Translation Adequacy}

Following the methodology outlined in \S\ref{sec:translation\_quality}, all post-edited output as well as all machine translations were evaluated by bilingual human judges using the adequacy scales shown in Table~\ref{judge\_guidelines} \vpageref{judge\_guidelines}.

\subsubsection{Rating Adequacy of Russian-English}

%\subsection{Rating of translations}

Following the adequacy guidelines from \S\ref{sec:translation\_quality}, %shown in Table~\ref{judge\_guidelines\_russian},

%

an experienced English-Russian translator and grader rated all English output translations of the Russian-English post-edited segments.

%

In addition, all English machine translations of the Russian documents were manually rated for adequacy.

% as well as the machine translation segments.

%in both human-readable tabular form and machine-readable CSV form in the attached dataset supplement.}

For each segment, the human rater was presented with a vertically-arranged list showing all variants of that segment.

%

The first entry in each list was the segment in the source language (Russian).

%

The source segment was followed by the reference translation in English.

%

The subsequent eight entries were English translations of the source segment, presented in a randomized order.

%

The English translations included the unedited machine translation output, as produced by Moses,

%

a post-edited translation produced by a monolingual post-editor from \citet{2014\_WMT\_Schwartz\_etal},

%

and the six post-edited translations produced by the Russian-English bilingual post-editors in this study.

All Russian-English translations were rated using the translation adequacy scale in Table \Vref{judge\_guidelines\_russian}, with possible ratings ranging from 2 (translation makes no sense) to 12 (translation is superior to the reference translation).

%The human rater was presented with blocks of 10 lines arranged in a vertical list:

%%

%the first line was a source segment,

%%

%the second was its reference translation, and

%%

%below these were eight translations for the segment

%%

%(the machine translation, one monolingual post-edited translation from \citet{2014\_WMT\_Schwartz\_etal}, and six bilingual post-edited translations)

%%

%listed in random order.

%%

%The rater was asked to classify all the segment translations on an adequacy scale (see Table \Vref{judge\_guidelines\_russian}) ranging from 2 (translation makes no sense) to 12 (translation is superior to the reference translation).

\subsubsection{Rating Adequacy of Spanish-English}

Following the adequacy guidelines from \S\ref{sec:translation\_quality},

%

an experienced Spanish-English translator and grader worked in cooperation with a second Spanish-English bilingual to rate all English output translations of the Spanish-English post-edited segments.

%

In addition, all English machine translations of the Spanish documents were manually rated for adequacy.

%

%One experienced Spanish-English translator and grader and one Spanish-English bilingual rated all post-edited segments as well as the MT segments.

%%

%They did this in cooperation.

For each segment, the human raters were presented with a vertically-arranged list showing all variants of that segment.

%

The first entry in each list was the segment in the source language (Spanish).

%

The subsequent twelve entries were English translations of the source segment, presented in a randomized order.

%

The English translations included the unedited machine translation output, as produced by Bing Translator,

%

and the eleven post-edited translations produced by the Spanish-English bilingual post-editors in this study.

Unlike the Russian documents, no reference translation was available for the Spanish documents; for this reason (as described in \S\ref{sec:translation\_quality}), the top category (12) used in evaluating Russian-English segments was omitted from the Spanish-English evaluation.

%

All Spanish-English translations were rated using the translation adequacy scale in Table \Vref{judge\_guidelines\_spanish}, with possible ratings ranging from 2 (translation makes no sense) to 10 (the meaning of the Spanish sentence is fully conveyed in the English translation).

%

Unlike all other participants, participant PE17 consistently produced post-edited segments of lower adequacy than the corresponding raw MT output. This participant was therefore dropped from all analyses of the Spanish-English data.

%

%

%%

%The raters were presented with blocks of twelve lines arranged in a vertical list: the first line was a source segment, and below these were the twelve translations for the segment listed in random order.

%%

%The raters were asked to classify all the segment translations on an adequacy scale (see Table \Vref{judge\_guidelines\_spanish}) ranging from 2 (translation makes no sense) to 10 (the meaning of the Spanish sentence is fully conveyed in the English translation).

%

%For the rating of the segments in this experiment the top category used for the evaluation of the Russian segments was removed because in the Spanish experiment we did not use a reference translation.

\subsection{Adequacy Results}

Figure~\ref{fig:percentage\_segments} \vpageref[above]{fig:percentage\_segments} presents the percentage of segments judged to be in each adequacy category.

%

Mean adequacy scores for each experimental condition are presented in Figure~\ref{fig:mean\_adequacy\_score} \vpageref[above]{fig:mean\_adequacy\_score}.

%

By subtracting the adequacy score of each machine translated segment, we obtain the adequacy gain obtained by post-editing;

%

these values are presented in Figure~\ref{fig:mean\_adequacy\_gain} \vpageref[above]{fig:mean\_adequacy\_gain}.

%The above effect can be seen clearly in Figure \Vref{fig:mean\_adequacy\_gain}, where we show mean gain in adequacy over unedited MT

%

%We also calculate the adequacy improvements of post-edited translations over unedited machine translations; such mean gains in adequacy are presented in Figure~\ref{fig:mean\_adequacy\_gain} \vpageref[above]{fig:mean\_adequacy\_gain}.

%

Finally, Figure~\ref{fig:mean\_adequacy\_score\_per\_posteditor} \vpageref[above]{fig:mean\_adequacy\_score\_per\_posteditor} presents mean adequacy score by post-editor.

%

We now analyze these results by experimental condition.

\subsubsection{Machine Translation Adequacy}

In Figure~\ref{fig:percentage\_segments} \vpageref[above]{fig:percentage\_segments}, we observe that the Russian-English machine translation segments tend to be of lower quality (as measured by adequacy), while the Spanish-English machine translation segments tend to be of higher quality.

%

Two-thirds of Russian-English machine translation segments are judged to have major errors (ratings 2-6), while one-third are rated as mostly or completely correct (8-12).

%

Contrast this with the Spanish-English machine translations segments;

%

a minority (about two-fifths) are judged to have major errors (ratings 2-6), while the majority (about three-fifths) are rated as mostly or completely correct (8-10).

\input{figure2}

\subsubsection{Russian-English Adequacy}

%The post-edited English segments

The mean adequacy score when bilingual participants were presented with alignments was 8.35.

%

When alignments were omitted from the post-editing tool, the mean adequacy score was 7.85.

%

A Wilcoxon signed-rank test \citep{1945\_Wilcoxon} showed that when participants were presented with alignments the ratings of their translations were significantly higher than when participants post-edited without access to alignments (N = 6, Z = -2.207, p = 0.027).

\subsubsection{Spanish-English Adequacy}

The mean adequacy score when Spanish bilingual participants were presented with alignments was 8.22. When alignments were omitted from the post-editing tool, the mean adequacy score was 8.02. These means are not significantly different.

%Our results suggest that when machine translation quality is low, bilingual post-editors may produce higher quality translations when presented with bilingual alignment links between source words and machine-translated target words.

%

%In addition to rating each post-edited translation, our bilingual human judge rated the adequacy of the unedited MT segments.

%%

%We thereby obtained direct human judgements of MT quality.

%%

%Because we care about the adequacy of post-edited translations, we consider actual human judgements to be preferable to automated metrics such as BLEU \citep{2002\_ACL\_Papineni\_etal}, which at best serve as a flawed proxy for human judgements.

%We partition the data according to the MT quality as measured by our human judge in terms of adequacy.

%%

%For each partition, we then calculate the mean adequacy score of post-edited segments, once for the case where post-editors were presented with alignment data, and once for the case where alignments were not presented.

%%

%In addition, we perform the same average calculation for the monolingual post-edited segments of Docs A and B from \citet{2014\_WMT\_Schwartz\_etal}.

%%

%The results are shown in Figure \ref{fig:mean\_adequacy\_score}.

%\clearpage

\input{figure3}

\subsection{Timing Results}

\subsubsection{Russian-English Timing}

The times taken by the Russian-English post-editors to post-edit each text were recorded manually to the nearest minute. The mean times were 33 minutes for texts with alignment and 40 minutes for texts without alignment. This difference approached significance (p = .082).

\subsubsection{Spanish-English Timing}

The times taken by the Spanish-English post-editors to post-edit each text were recorded by the keylogger to nearest millisecond. The mean times were 21 minutes and 5 seconds for texts with alignment and 22 minutes and 5 seconds for texts without alignment. An independent samples t-test showed that these times are not significantly different from each other (t(18) = .295, p = .77.)

%

Participants were not given time limitations, so timing data must be interpreted with caution. Note, however, that the mean time with alignment is numerically shorter than the mean time without alignment.

The shorter editing times for Spanish-English may in part be explained by the shorter length of these documents.

\section{Analysis and Related Work}

\label{sec:background}

Our results suggest that when machine translation quality is poor (2--4), bilingual post-editors may produce higher quality translations when presented with bilingual alignment links between source words and machine-translated target words.

%

Alternatively, when machine translation quality is high (8--12), no effect is seen by presenting alignment visualizations.

%

We explain this by hypothesizing that word alignment visualization may enable post-editors to better recover from certain types of translation errors produced by MT systems; when MT quality is high enough that such errors are absent, word alignment visualization may no longer play a restorative role.

%

%

%We observe that when MT quality was poor (adequacy categories 2--6), bilingual post-editors consistently produced higher quality post-edited translations when alignments were shown.

%

%On the other hand, when MT quality is high (adequacy categories 8--10), little or no impact to adequacy quality was observed when alignments were shown.

%%

%By subtracting the adequacy score of each machine translated segment, we obtain the adequacy gain obtained by post-editing.

%%

%The above effect can be seen clearly in Figure \Vref{fig:mean\_adequacy\_gain}, where we show mean gain in adequacy over unedited MT;

%%

%post-editing makes the most difference when MT quality is poor, and in those cases the presence of alignment links improves post-editing as measured by adequacy.

We examine the effect that alignment link visualization has on each bilingual post-editor in Figure \ref{fig:mean\_adequacy\_score\_per\_posteditor} \vpageref[above]{fig:mean\_adequacy\_score\_per\_posteditor}.

%

In the Russian-English condition, where overall MT quality is poor, we observe that post-editing quality varies widely between post-editors (with PE2 and PE3 performing best).

%

For all six bilingual post-editors, we observe higher mean adequacy scores when alignment links were presented than when they were omitted from the post-editing tool.

%

We also note that when alignment links were absent, one bilingual post-editor (PE5) performed worse than the monolingual post-editor (PE0) from \citet{2014\_WMT\_Schwartz\_etal}.

%

On the other hand, in the Spanish-English condition, where overall MT quality is good, we observe relatively little variation in quality between the ten post-editors.

%

When compared to the unedited machine translations, post-editing resulted in improved mean adequacy for all post-editors, both bilingual and monolingual.

Our results also suggest that post-editing time tends to be reduced for texts with alignment.

%

These numerical reductions were consistent across all the Russian-English participants with the exception of PE3, but they were not consistent for the Spanish-English participants.

We hypothesize that texts with alignment are less cognitively demanding to process, and so less effortful to post-edit than texts without alignment.

%

If this is the case, shorter post-editing times for texts with alignment are consistent with previous findings by \citet{2012\_AMTA\_Koponen\_etal}, who found that per word post-editing times were shorter for segments that were less cognitively demanding because of the linguistic structure.

%

Related work on cognitive effort in post-editing \citep{2014\_AMTA\_Lacruz\_etal,2014\_Lacruz\_Shreve} has also shown decreased densities of short pauses when less cognitively demanding segments are post-edited.

The keystroke logging data gathered for Spanish-English post-editors allowed the computation of Pause to Word Ratio (PWR). For each segment, PWR is the ratio of the number of pauses exceeding 300ms to the number of words in the MT segment; it is a measure of cognitive effort in post-editing (Lacruz & Shreve, 2012). Higher PWR corresponds to higher cognitive effort.

% INSERT REF TO FIG 6 AND INSERT FIG 6

Contrary to expectation, the mean PWR for Spanish-English post-editors was slightly higher for the segments with alignment (0.70) than for those without alignment (0.63). However, the numerical difference was not significant.

It is possible that the effect of alignment on PWR was masked by the fact that the adequacy of the Spanish-English MT segments was generally high. Since our prediction is that alignment should both increase post-editing adequacy and reduce post-editing effort, the Gain to Effort Ratio, GER = (PE Rating – MT Rating)/PWR is a promising metric to investigate. We hypothesize that GER is higher for segments with alignment than for segments without alignment.

GER for segments with alignment was 2.53, and GER for segments without alignment was1.47. Our prediction was confirmed: a paired samples t-test showed that GER was higher for segments with alignment (t(9) = 2.49, p =.034.)

%INSERT REF TO FIG 7 AND INSERT FIG 7

This result suggests that GER may be a robust metric for measuring the effects of alignment on post-editing. It would be interesting to conduct further studies involving language pairs different from Spanish-English where the adequacy of machine translations may be lower.

%Finally, in Figure \ref{fig:percentage\_segments} we present histograms graphing the percentage of segments in each adequacy category.

%%

%We observe that the majority of unedited machine translations are of relatively poor quality, falling mostly into categories 4 and 6. %, and a few scoring as 12.

%%

%%The monolingual post-editor from \citet{2014\_WMT\_Schwartz\_etal} was able to skew the quality histogram somewhat to the right, with most segments scored as 6 or 10, and a few scored as better than the reference (12).

%The monolingual post-editor improved MT quality somewhat, with most segments scoring as 6 or 10.

%%

%The histogram for bilingual post-editors using no alignments skewed further to the right, peaking at 8.

%%

%The histogram for bilingual post-editors with access to alignments skews farthest to the right, peaking at 10.