We should try to use consistent terms for translation/hypothesis/system output

How do humans evaluate Machine Translation

First Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Second Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

2) it is complex.. Judges can have difficulty distinguishing between different aspects of the translations, specially if guidelines are not well defined

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam at purus vel tellus tincidunt tristique et a tellus. Donec sagittis, dui et efficitur maximus, enim felis gravida libero, eget varius erat quam in nibh. Aenean nec libero odio. Praesent posuere, nisl a eleifend fermentum, nulla lorem congue ante, id mollis velit lectus id ex. In dapibus non arcu ac sollicitudin. Etiam commodo erat nec sapien bibendum facilisis. Praesent porttitor at magna quis blandit. Proin mattis pellentesque justo non hendrerit. Suspendisse potenti. In hac habitasse platea dictumst.

Praesent congue nulla a accumsan venenatis. Maecenas at nisl et mi congue placerat quis in dolor. Cras consectetur lacus id dui commodo posuere. Sed sit amet urna tincidunt, facilisis sem eu, feugiat nibh. Ut eu dui non nisl feugiat congue. Ut sit amet turpis sed justo venenatis ornare suscipit vel sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam sollicitudin metus nec nunc pellentesque convallis. Morbi sit amet lectus ac mauris hendrerit luctus sed in justo.

1 Introduction

I did not get the first sentence. we may remove from "and decide"

Each year thousands of human-judgments are used to evaluate the quality of Machine Translation (MT) systems, ~~and decide which algorithms and training techniques are to be considered the new state of the art.~~ In ~~the~~ a typical scenario human judges evaluate a system's output (or *hypothesis*) by comparing it to a source sentence or /and a reference translation. Then, they score the hypothesis according to a set of defined criteria such as *fluency* and *adequacy* (White et al., 1994); or rank ~~a~~ it against a set of hypotheses in order of preference (Vilar et al., 2007; Callison-Burch et al., 2007).

They dont compare with a source sentence, instead they use it

The hypothesis is scored according to a ...

Evaluating MT output can be a challenging task for two main reasons:1) it is tedious and evaluators can lose interest quickly; ~~and~~2)it is complex, specially if ~~the~~ guidelines are not well defined, ~~and~~ judges can have difficulty distinguishing between different aspects of the translations (Callison-Burch et al., 2007). Furthermore, it is highly subjective as users develop their own *intrinsic* rules of thumb. As a result, evaluations suffer from low inter- and intra-annotator agreements (Turian et al., 2003; Snover et al., 2006). Yet, as Sanders et al. (2011) argue, using human judgments is essential to the progress of MT because: (*i*) automatic translations are produced for a human audience; and (*ii*) human understanding of the *real* world allows to assess the importance of the errors made by MT systems.

can we rearrange it. See a suggestion above

we can remove this last sentence. we are talking about evaluation issues and then suddenly we jump to the importance of human evaluation

I did not get point ii at all

Most of the research in human evaluation has focused on analyzing the criteria to use for evaluation, and has regarded the evaluation process as a *black-box*, where the input are different sources of information (i.e source text, reference translation, and translation hypotheses), and the output is a score (or preference ranking).

is there a downside of using it as blackbox?

citation?

In this paper, we rather focus on analyzing evaluation from a different perspective. First we regard the process as a *glass-box*, and use eye-tracking to monitor the times users spend digesting different sources of information before making a judgment. Secondly, we contrast how the availability of such sources can change the outcome of the evaluation. Finally, we try to characterize the background of the evaluator (in this case whether they are *monolingual* or *bilingual*) as a possible source of bias.

Our main research questions are:

what is current layout? it is not introduce before. Given a layout?

- Given the current layout, what kind of information do evaluators prefer to use to decide on the score of a ~~translation~~ hypothesis:

Should we say something about the appraisal code that we are making available to facilitate research?

The description of Callison work is too long and sometimes redundant. We should summarize it. We need to relate it to what we are dong rather than completely describing their work.

  - Do they use the source text, the target text, or both?
  - Do they follow a specific order of areas of interest while evaluating?
  - How long time do they spend in each area and in general to give an evaluation score to a translation?

Suggestion to make below point sound diff from last point: are bilinguals different from monolinguals in evaluation?

- Are there differences of behavior between *bilinguals* (i.e. subjects fluent in both source and target languages) and *monolinguals* (i.e. subjects fluent only in the target language) ? ~~Which group is more consistent?~~

- Are there differences in behavior when evaluating *good* vs. *bad* translations?

the below sentence can be move above right before presenting research questions

Our goal is to provide actionable insights that can help to improve the process of evaluation, especially in large-scale shared-tasks such as WMT.

The remainder of this paper is divided as follows: First, we give a ...

need some work in related work section

2 Related Work 1

what do we mean by scenarios below?

~~Different authors have focused on assessing different aspects of the human evaluation process. From a categorization of the possible scenarios (Sanders et al., 2011) to the effectiveness of the evaluation criteria (Callison-Burch et al., 2007).~~

~~In the past,~~ researchers have proposed different methods to assess the quality of a translation, such as the direct evaluation of *adequacy* and *fluency*, and the ranking-based evaluation (Vilar et al., 2007; Callison-Burch et al., 2007). Unfortunately, humans have a hard time assigning an absolute score to a translation, and in major MT evaluations, absolute scores were phased out in favor of ranking-based evaluations or task-based evaluations (e.g. HTER). It has been shown that using such ranking-based assessments yields much higher inter-annotator agreement (Callison-Burch et al., 2007). However, even in ranking-based tasks, annotators have different criteria of what makes a *good* or a *bad* translation. This ~~is often~~ can be ? determined by their background or level of experience and can ~~determine~~ be used to evalute the performance of an evaluator, as well as undermine the quality of the evaluation.

According to Callison-Burch et al. (2007), there are several criteria that define the MT evaluation task: (*i*) The *ease* with which humans are able to perform the task; (*ii*) the agreement w.r.t. other annotators, and the *speed* with which annotations can be collected. They assessed three different ways of evaluating sentences: 1. Assign~~ing~~ scores according to a five-point scale for adequacy and fluency~~, following the guidelines described in (LDC, 2005),~~ 2. Rank~~ing~~ (five) translated sentences relative to each other's quality from best to worst. 3. Rank~~ing~~ the translations of selected *syntactic constituents* drawn from the source sentence.

~~According to Callison-Burch et al. (2007), there are several criteria that define the MT evaluation task: (*i*) The *ease* with which humans are able to perform the task; (*ii*) the agreement w.r.t. other annotators, and the *speed* with which annotations can be collected.~~

In their observations, people had a hard time distinguishing between the *fluency* and *adequacy* aspects of a translation, and found that there is a high correlation between both scores. ~~Furthermore,~~ the lack of clear guidelines further complicates the assessment (e.g. how is *meaning* quantified, how do grammatical errors affect different levels of fluency). Therefore, Callison-Burch et al. (2007) point out that each annotator develops their own rules of thumb. By asking annotators to rank the hypotheses, the task was simplified considerably, and inter-annotator agreement increased. Furthermore, when asked to rank only constituents, additional improvements were observed. In terms of time, ranking constituents reduced the evaluation time to about 11 seconds on average, when compared to 26 seconds on average for the full hypothesis. Note however, that the times were not normalized by the number of words in the hypothesis/constituent. In summary, Callison-Burch et al. (2007) recommended that evaluations are carried with ranking experiments instead of fluency/adequacy absolute scores. Yet, specific criteria to consider when evaluating a translation were not provided. Instead, instructions are "kept minimal" by only asking the evaluator to rank hypotheses from worst to best (Bojar et al., 2011).

However, we rarely question our assumptions about the task. ???

[FG]:The motivation behind using eye-tracking is to extract information on how users evaluate a translation. Intuitively, we would like to use that information to reduce the bias in the process of evaluating a translation.

As mentioned before, there are two main challenges when presenting an evaluation task: (*i*) how to make the task *less* tedious, i.e. increasing engagement, to preserve the user's focus; and (*ii*) how to develop consistent guidelines to increase inter-annotator agreement. For (*i*) , (Doherty et al., 2010) proposed to have a comprehension questionnaire aimed to encourage the user's focus retention, while for (*ii*) , ranking tasks have been proposed. Instead, here we propose a evaluation as a game, in which users have to build their own strategies to mimic the evaluation of an *expert* translator. This serves to both purposes: to keep the users engaged by immediately given feedback after each evaluation; and to train users to develop consistent strategies that mimic a *gold standard*.

I

3 Method

3.1 Data

Here we talk about the WMT2012 data we used. why did we choose Spanish-English?

We also provide statistics. How many users evaluated each sentence?

We decided to keep only a fraction of the data. Explain the reasoning behind.

How did we group the source sentences? Short, mid long? How was this split determined?

How did we choose the translations. Explain the rationale. We wanted to contrast evaluation of good vs bad translations.

3.2 Layouts and Areas of Interest

We choose Appraise an open-source toolkit [1] for translation evaluation to conduct our experiments. Typical Appraise evaluation settings provides the user with (*i*) the hypothesis (or hypotheses) to be evaluated; (*ii*) the source sentence; (*iii*) the context of the source sentence (previous and next sentences in the same document); (*iv*) the reference translation for the source sentence; (*v*) the context of the reference translation. Additionally, we added a slider for the user to provide a score for the quality of translation and a feedback to the user after submitting his evaluation. The user was given a feedback about his scoring in reference to the original judgment of the task sentence (See Figure 1.

To exclude the effect of the layout, we explored using three different evaluation scenarios -games-:

[1] Available at:github.com/cfedermann/Appraise

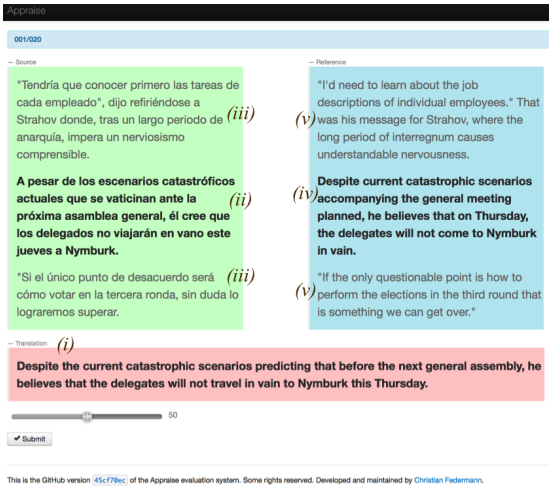

Figure 1: Appraise evaluation layout

- **Scenario 1** shows participants the translated sentence ~~(in English)~~ along with the source text of the translation ~~(in Spanish),~~ including the context (1 sentence before and 1 sentence after the translated sentence).

- **Scenario 2** shows participants the translated sentence ~~(in English),~~ along with the source text of the translation ~~(in Spanish),~~ and a reference translation done by a human ~~(also in English).~~

- **Scenario 3** shows the translated sentence ~~(in English)~~ only with a reference translation done by a human ~~(in English).~~

The source and the reference translation include the context of the sentence, i.e. 1 sentence before and 1 sentence after the sentence.

We first conduct multiple pilot experiments involving 4-5 users \footnote{This includes the author of this article}.

3.3 Experiment Participants

The actual experiment was preceded by multiple rounds of pilot experiments involving 4-5 users, including the authors of this article. This step was necessary in order to fine tune the eye-tracker and exclude any external influences such as light embedding the correct recording of the eyes or noise, distracting the participants.

In the actual experiment participated 21 users aged 27 to 45 years old. restructure this sentence Most of the participants (16) were computer scientists. Out of them, 9 had experience with manually translating documents, and 4 had experience with machine translation evaluation. 7 of the participants were female, and 13 were male.

Information regarding the use of glasses/lenses (and whether they are anti-reflective) during the experiment was collected. This information was necessary to rule out any interference with the eye-tracker due to the use of glasses or lenses. Out of the 20 users, 6 weared glasses (5 were anti-reflective), and 2 users weared lenses. Only in the case of one participant, there was interference with the eye-tracker, which could not correctly track his/her eye movements, so the participant was not taken into account.

The participants, who were good for the experiment were divided into two groups, each composed of 10 people:

- **Bilingual** participants did speak the source language of the translation (Spanish) at satisfactory level of comprehension (native or advanced).
- **Monolingual** participants did not speak the source language of the translation.

The participants spoke a variety of other languages (besides the target language English), including: Arabic, Turkish, German, Danish, Bulgarian, Russian, Slovenian, Croatian, Hindi, Chinese, Basque. The number of languages per person varied between 1 and 9 languages.

A shortcoming of the experiment design was that since Spanish is a quite expanded language, our hypothesis was that the speakers of close Romance languages (such as Italian, French, or Portuguese) could also understand the source language to a certain point. For this reason, an extensive information regarding the languages participants mastered, and to what extent, was conducted. The speakers of other Romance languages, however, insisted that their knowledge of another romance language was not enough to correctly comprehend the source text in Spanish, so we ruled out this hypothesis.

Out of the 10 monolingual users, 6 spoke other Romance languages, namely:

- 1 user spoke beginner Italian
- 1 user spoke native Italian
- 1 user spoke native Italian and beginner French
- 2 users spoke beginner French
- 1 user spoke native Portuguese

3.4 Experimental Design

How was the experiment designed

What were the variables considered?

- User type
- Translation type
- Length type
- Game type (layout)
- repetitions

Explain the design matrix used. What is the total number of evaluations collected in this round? 1200, 60 per user.

3.5 Eye-tracking Setup

To accommodate our experiments, we augmented Appraise Quality Estimation task to incorporate information from eye tracker. The new "Eye-Appraise" includes the capability to interface with Eyetracker and record eye information along with the user judgments. We used the low cost "Eye-Tribe" (Eyetribe, 2014) eye tracker to collect eye information and communicates the data to the Tracker server which serve to broadcast the data/messages read by the tracking device to Appraise. The message received by Eye-Appraise contains information about the position of both left and right eyes, fixation, time stamp and a state indicating the status of the tracker. Eyetribe user guide [2] details further the structure and the attributes of the messages that can be exchanged with the server. Figure 2. details the architecture of "Eye-Appraise". The Eyetribe device was set to operate at a frequency of 60Hz (a reading in every 16ms) which makes it not suitable to track high accuracy saccades.

3.6 Instructions and Exit Poll

What were the instructions given to the users? Why?

What were the considered variables? [FG]:@Irina

4 Results

The data gathered from the eyetracker revealed valuable insights about evaluators behavior and their approach to evaluation. The data collected showed that background of users has an impact on

[2] http://dev.theeyetribe.com/api/

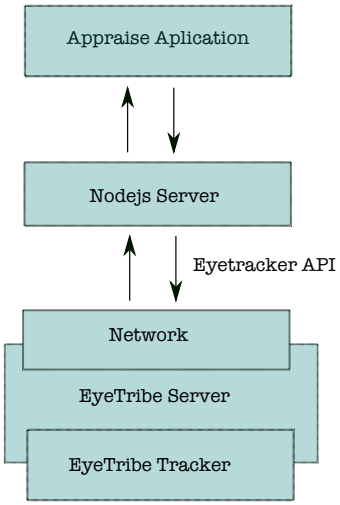

Figure 2: Eyetribe communication map

translation. While both users type (mono- and bilinguals) uses all the information available to them as shown in Figure 3 and 4; both they tend to use the information is different way depending on the game and quality. In the next sections, we will dive deeper on the various aspects of these differences trying to answer our research questions and reveal key information about the evaluation process.

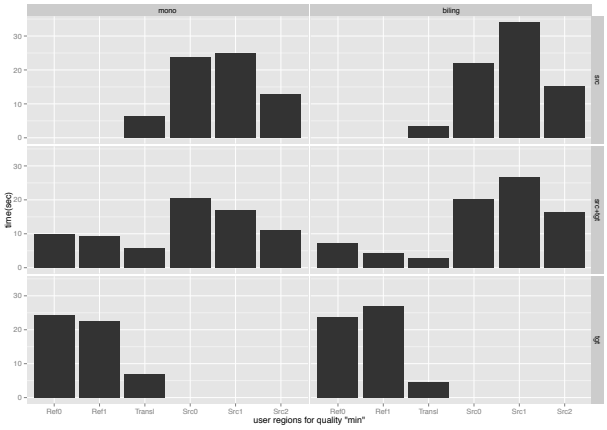

Figure 3: Users time in regions while scoring "min" category.

4.1 Time as a Criterion

Analyzing user's speed from Figures 5 and **??**, shows that bilingual users are relatively faster in completing tasks for both max and min types. Bilingual mean, max is (27.51,64.91) versus (33.78,121.65) for the Monolinguals. Monolingual user's speed varies widely. Scoring "max" category takes more time for both user types. Here

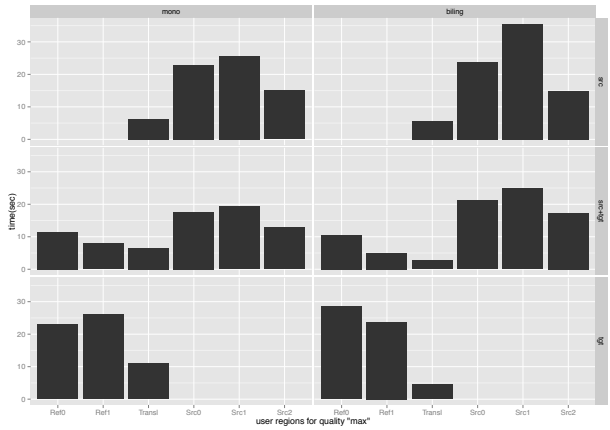

Figure 4: Users time in regions while scoring "max" category.

we evaluate how users's speed is, and how does it affect results if any.

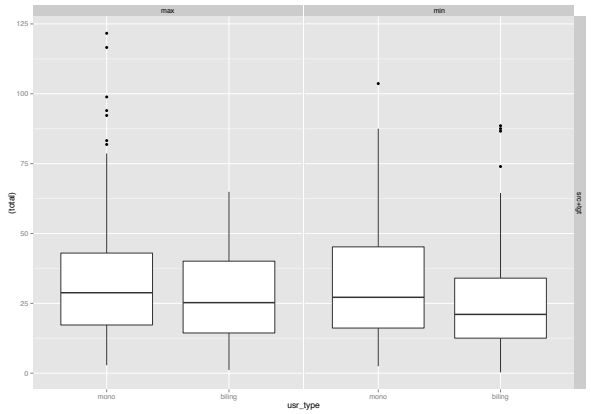

Figure 5: Users time in src+tgt scenario

4.2 Score Consistency

Here we tell how users are consistent with each other and with the ranking/provided scores

4.3 User Behavior

Figures 7 and 8 present the users behavior while evaluating sentences from "max" category. The circles size represents the proportional time spent in the area while the directed arcs width reflect the frequency of the transition between the connected regions. From the figures, we can note that the bilingual users tend to spend more time between source -See Figure 7- regions when compared to monolinguals. Monolinguals move back and forth between source and reference more frequently when scoring "max" category as shows Figure 8.

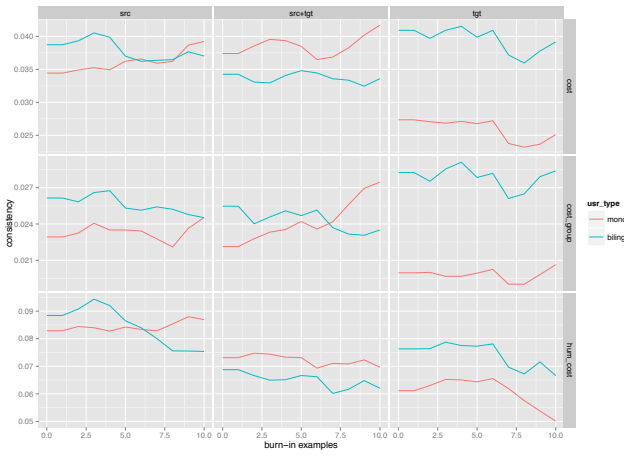

Figure 6: Users consistency in various scenarios

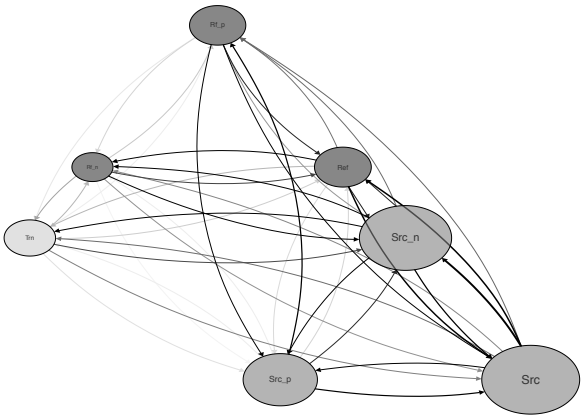

Figure 8: Monolingual user process of scoring "max" category.

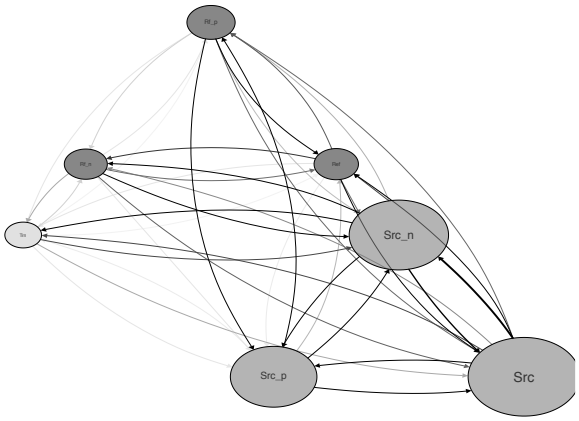

Figure 7: Bilingual user process of scoring "max" category.

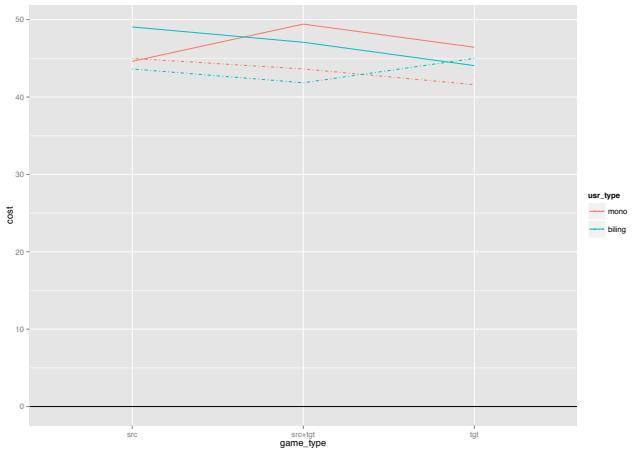

Figure 9: Consistency of users vs humans and language model

Conclusions:
The results shows that while bilingual evaluators are typically faster that their monolingual peers, the latters are more consistent in their judgments. Monolingual inability to access source data, it is not an obstacle for them to make an accurate judgment. On the other hand, when the hypotheses were scored with a language model. Monolingual looks like they are more consistent in mimicking language model scores as shown in Figure 9 - Dashed lines.

5 Discussion

5.1 Why do we need feedback?

5.2 Can we do more with eye-tracking?

Eye-tracking technology has proven useful in different scenarios related to translation. [FG]:Here we give more examples of how it was used

So far, we have used the eye-tracking device to measure the *dwell* time a user spends reading a specific portion of the screen. Nonetheless, one can think of more refined uses for this technology.

Potentially, using eye-tracking can give us a fine-grained insight on how users differentiate *good* from *bad* translations, making it easier to *learn* the intrinsic rules of thumb that they use during the evaluation process. The applications for this are manifold. For example, by learning which type of errors (e.g. morphological, syntactic, semantic) produce more disturbances in the expected evaluation behavior, we could help to develop *better* automatic MT evaluation metrics. Additionally, we can use gaze-data to model the evaluation score (or rank) given by a user, and thus reduce the subjective score bias. This can help to alleviate the high variance found in evaluation.

However, there are several challenges that need to be solved before moving forward in this nascent area. The most important is related to the accuracy of the eye-tracking devices, which is a requirement to track which specific words are looked-at in the screen. Eye-tracking errors can be divided into two categories: variable (device precision) and systematic. Fortunately, the former has improved over the past years, and high-accuracy devices can be now acquired for only a few hundred dollars. The latter, however is more complex. Often, a loss in accuracy known as *drift* is observed as time progresses, requiring frequent re-calibrations of the eye-tracking device. This can be due to user movements, and other environmental factors. Reducing and eliminating drift is imperative to make progress in this area. Up to now, only heuristic approaches have been proposed (Mishra et al., 2012).

5.3 Is bilingual adequacy necessary?

6 Related Work

Eye-tracking has been previously used also in MT Evaluation research (Stymne et al., 2012; Carl, 2012; Alabau et al., 2014; Doherty and O'Brien, 2014; Doherty et al., 2010). The papers more relevant to us are Doherty et al. (Doherty et al., 2010) and Stymne et al. (2012).

The paper closest to us is Doherty et al. (Doherty et al., 2010). They have used the Tobii 1750 eye-tracker. 10 native speakers of French were asked to read and evaluate for comprehensibility 50 sentences, translated by an automatic machine translation system into French. 25 of the sentences were rated as excellent in previous human rating, and 25 - as poor. Four eye-tracking variables were used: 1) gaze time; 2) fixations count; 3) pupil dilation; and 4) average fixation duration.

Stymne et al. (2012) applied eye-tracking to machine translation error analysis. 33 university students were asked to read the text outputs of three machine translation systems, along with a human-produced text, and evaluate the text quality. Eye-tracking was recorded using SMI Remote Eye iView, and the following variables were recorded: 1) average gaze time and 2) fixations count.

Doherty and O'Brien (2014) use eye-tracking to evaluate the quality of raw machine translation output, in terms of its "usability" by an end user. In order to achieve this, an online service documentation was translated using Google Translate, and 30 participants were asked to read the documentation, and perform tasks, based on this documentation. An eye-tracking tool, Tobii 1750, was used to record the eye-movements of the participants while reading the documentation and executing the tasks in order to measure the cognitive effort involved in processing the documentation. The following eye-tracking variables were used: 1) fixations count; and 2) average fixation duration.

In addition to the aforementioned works, Translog-II (Carl, 2012) and Casmacat (Alabau et al., 2014) are two (computer-aided) translation workbenches, which allow as an option the integration of an eye-tracker (such as Tobii). The eye-tracking in Translog-II records 1) gaze-sample points for left and right eyes; 2) fixations.

7 Conclusion

Sed sem massa, feugiat non dui eget, faucibus imperdiet lacus. In aliquet et lacus non blandit. In mi nisl, auctor ut accumsan id, placerat ut ante. Donec augue nisl, venenatis at nibh ac, lacinia imperdiet ligula. Morbi hendrerit gravida felis, sed malesuada magna auctor sit amet. Donec sit amet lobortis nibh. Duis eleifend justo vitae justo faucibus pellentesque. In non nunc quis risus mattis interdum. Praesent blandit vitae tellus a eleifend.

Mauris porta tellus eu leo bibendum ultricies. Pellentesque eget sem ac ante interdum tincidunt porta ut diam. Donec malesuada dolor massa, ut laoreet turpis ullamcorper non. Nunc tincidunt lacus ac nisl cursus, in feugiat nunc imperdiet. Nam consequat tincidunt augue. Integer ut volutpat nisi. Aenean id lorem quam. Vestibulum vulputate, lectus vel porta eleifend, nibh odio gravida est, sit amet posuere urna nisl pulvinar sapien. Phasellus luctus massa nulla. Donec sed pharetra erat. Duis ac ipsum tincidunt, eleifend augue ut, pharetra turpis.

Quisque placerat est vel magna volutpat malesuada. Proin feugiat, libero sed egestas vestibulum, ipsum justo faucibus diam, commodo faucibus sem quam at leo. Donec ut tellus at lorem consequat rhoncus. Duis ac maximus ante, sed porta lorem. Sed pellentesque ultrices. (Bojar et al., 2011)

References

Vicent Alabau, Christian Buck, Michael Carl, Francisco Casacuberta, M Garcıa-Martınez, Ulrich Germann, Jesús González-Rubio, Robin Hill, Philipp Koehn, LA Leiva, et al. 2014. Casmacat: A computer-assisted translation workbench. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 25–28.

Ondrej Bojar, Miloš Ercegovčević, Martin Popel, and Omar Zaidan, 2011. *Proceedings of the Sixth Workshop on Statistical Machine Translation*, chapter A Grain of Salt for the WMT Manual Evaluation, pages 1–11. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, WMT '07, pages 136–158, Prague, Czech Republic.

Michael Carl. 2012. Translog-II: a program for recording user activity data for empirical reading and writing research. In *LREC*, pages 4108–4112.

Stephen Doherty and Sharon O'Brien. 2014. Assessing the usability of raw machine translated output: A user-centered study using eye tracking. *International Journal of Human-Computer Interaction*, 30(1):40–51.

Stephen Doherty, Sharon O?Brien, and Michael Carl. 2010. Eye tracking as an MT evaluation technique. *Machine translation*, 24(1):1–13.

Eyetribe. 2014. Getting started: Setting up the eyetribe. retrieved from http://dev.theeyetribe.com/start/ on mar 31, 2015.

LDC. 2005. Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Technical report, Technical report.

Abhijit Mishra, Michael Carl, and Pushpak Bhattacharyya. 2012. A heuristic-based approach for systematic error correction of gaze data for reading. In *Proceedings of the First Workshop on Eye-tracking and Natural Language Processing*, pages 71–80, Mumbai, India, December. The COLING 2012 Organizing Committee.

Gregory Sanders, Mark Przybocki, Nitin Madnani, and Matthew Snover. 2011. Human subjective judgments. In Joseph Olive, Caitlin Christianson, and John McCary, editors, *Handbook of Natural Language Processing and Machine Translation*, pages 806–814. Springer.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas*, AMTA '06, Cambridge, Massachusetts, USA.

Sara Stymne, Henrik Danielsson, Sofia Bremin, Hongzhan Hu, Johanna Karlsson, Anna Prytz Lillkull, and Martin Wester. 2012. Eye tracking as a tool for machine translation error analysis. In *LREC*, pages 1121–1126.

Joseph Turian, Luke Shen, and I. Dan Melamed. 2003. Evaluation of machine translation and its evaluation. In *In Proceedings of MT Summit IX*, pages 386–393, New Orleans, LA, USA, September.

David Vilar, Gregor Leusch, Hermann Ney, and Rafael E. Banchs. 2007. Human evaluation of machine translation through binary system comparisons. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 96–103, Prague, Czech Republic, June. Association for Computational Linguistics.

John White, Theresa O'Connell, and Francis O'Mara. 1994. The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. In *Proceedings of the Association for Machine Translation in the Americas conference*, pages 193–205, Columbia, Maryland, USA, October.