Eyes Don't Lie: Predicting Machine Translation Quality Using Eye Movement

Hassan Sajjad, Francisco Guzmán, Nadir Durrani, Ahmed Abdelali,
Houda Bouamor†, Irina Temnikova, Stephan Vogel
Qatar Computing Research Institute, HBKU, Qatar
†Carnegie Mellon University, Qatar

Abstract

Poorly translated text is often disfluent and difficult to read. In contrast, well-formed translations require less time to process. In this paper, we model the differences in reading patterns of Machine Translation (MT) evaluators using novel features extracted from their gaze data, and we learn to predict the quality scores given by those evaluators. We test our predictions in a pairwise ranking scenario, measuring Kendall's tau correlation with the judgments. We show that our features provide information beyond fluency, and can be combined with BLEU for better predictions. Furthermore, our results show that reading patterns can be used to build *semi*-automatic metrics that anticipate the scores given by the evaluators.

1 Introduction

Human evaluation has been the preferred method for tracking the progress of MT systems. In the past, the prevalent criterion was to judge the quality of a translation in terms of *fluency and adequacy*, on an absolute scale (White et al., 1994). However, different evaluators focused on different aspects of the translations, which increased the subjectivity of their judgments. As a result, evaluations suffered from low inter- and intra-annotator agreements (Turian et al., 2003; Snover et al., 2006). This caused a shift towards a ranking-based approach (Callison-Burch et al., 2007). Unfortunately, the disagreement between evaluators is still a challenge that cannot be easily resolved due to the non-transparent thought-process that evaluators follow to make a judgment.

The eye-mind hypothesis (Just and Carpenter, 1980; Potter, 1983) states that when completing a task, people cognitively process objects that are in front of their eyes (i.e. where they fixate their gaze).[1] Based on this assumption, it has been possible to study reading behavior and patterns (Rayner, 1998; Garrod, 2006; Hansen and Ji, 2010).

The overall difficulty of a sentence and its syntactic complexity affects reading behavior (Coco and Keller, 2015). Ill-formed sentences take longer to process, and may cause the reader to jump back while reading. Hence, by looking into how evaluators read the translations and their accompanying references, we can learn about: (*i*) the complexity of a reference sentence, and (*ii*) the quality of a translation sentence.

Using reading patterns from evaluators could be a useful tool for MT evaluation: (*i*) to shed light into the evaluation process: e.g. the general reading behavior that evaluators follow to complete their task; (*ii*) to understand which parts of a translation are more difficult for the annotator; and (*iii*) to develop *semi*-automatic evaluation systems that use reading patterns to predict translation quality.

In this paper, we make a first step towards (*iii*): using reading patterns as a method for distinguishing between good and bad translations. Our hypothesis is that bad translations are difficult to read, which may be reflected by the reading patterns of the evaluators. Motivated by the notion of reading difficulty, we extracted novel features from the evaluator's gaze data, and used them to model and predict the quality of translations as perceived by evaluators.

[1] Except in cases of *covert* attention.

Proceedings of NAACL-HLT 2016, pages 1082–1088,
San Diego, California, June 12-17, 2016.

2 Features and Model

A perfectly *grammatical* sentence can be difficult to read for several reasons: unfamiliar vocabulary, complex syntactic structure, syntactic or semantic ambiguity, etc. (Harley, 2013). Reading automatic translations is even more challenging due to untranslated words, incorrect word order, morphological disagreements, etc. Cognitively processing difficult sentences generally results in modified reading patterns (Garrod, 2006; Coco and Keller, 2015).

In this paper, we analyze the reading patterns of human judges in terms of the word transitions (jumps), and the time spent on each word (dwell time); and use them as features to predict the quality score of a specific translation. For the sake of simplicity, as recommended by Guzmán et al. (2015), we only consider a monolingual evaluation scenario and ignore the source text . However, our features and experimental setup can be extended to include source-side features.

2.1 Features

Jump features While reading text, the gaze of a person does not visit every single word, but it advances in jumps called *saccades*. These jumps can go forwards (*progressions*) or backwards (*regressions*). The number of regressions correlates with the reading difficulty of a sentence (Garrod, 2006; Schotter et al., 2014; Metzner, 2015). In an evaluation scenario, a fluent reading would mean monotonic gaze movement. On the contrary, the reader may need to jump back multiple times while reading a poor translation. We classify the word-transitions according to the direction of the jump and distance between the start and end words. For subsequent words n, $n+1$, this would mean a forward jump of distance equal to 1. All jumps with distance greater than 4 were sorted into a 5+ bucket. Additionally, we separate the features for reference and translation jumps. We also count the total number of jumps.

Total jump distance We additionally aggregate jump distances[2] to count the total distance covered while evaluating a sentence. We have reference distance and translation distance features. Again, the idea is that for a well-formed sentence, gaze distance should be less, compared to a poorly-formed one.

Inter-region jumps While reading a translation, evaluators can jump between the translation and a reference to compare them. Intuitively, more jumps of this type could signify that the translation is harder to evaluate. Here we count the number of transitions between reference and translation.

Dwell time The amount of time a person fixates on a region is a crucial marker for processing difficulty in sentence comprehension (Clifton et al., 2007) and moderately correlates with the quality of a translation (Doherty et al., 2010). Our feature counts the time spent by the reader on each particular word. We separate reference and translation features.

Lexicalized Features The features discussed above do not associate gaze movements with the words being read. We believe that this information can be critical to judge the overall difficulty of the reference sentence, and to evaluate which translation fragments are problematic to the reader. To compute the lexicalized features, we extract streams of reference and translation lexical sequences based on the gaze jumps, and score them using a tri-gram language model. Let $R_i = r_1, r_2, \ldots, r_m$ be a sub-sequence of gaze movement over reference and there are $R_1, R_2, \ldots, R_n$ sequences, the lex feature is computed as follows:

$$lex(R) = \sum_{i}^{n} \frac{\log p(R_i)}{|R_i|}$$

$$p(R_i) = \sum_{j}^{m} p(r_j | r_{j-1}, r_{j-2})$$

The normalization factor $|R_i|$ is used to make the probabilities comparable. We also use unnormalized scores as additional feature. A similar set of features $lex(T)$ is computed for the translations. All features are normalized by the length of the sentence.

2.2 Model

For predicting the quality scores given by an evaluator, we use a linear regression model with ridge

[2] Jump count and distance features have also shown to be useful in SMT decoders (Durrani et al., 2011).

regularization. The ridge coefficient $\hat{\beta}$ is the value of β that minimizes the error:

$$\sum_i (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Here the parameter λ controls the amount of shrink applied to regression coefficients. A high value of λ shrinks the coefficients close to zero (Hastie et al., 2001). We used the implementation provided in the glmnet package of R (Friedman et al., 2010), which inherits a cross-validation mechanism that finds the best value of λ on the training data.

3 Experimental Setup

We used a subset of the Spanish-English portion of the WMT'12 Evaluation task. We selected 60 medium-length sentences which have been evaluated previously by at least 2 different annotators. For each sentence we selected the *best* and *worst* translations according to a human evaluation score based on the *expected wins* (Callison-Burch et al., 2012). As a result, we had 60 references with two corresponding translations each, adding up to a total of 120 evaluation tasks. Each evaluation task was performed by 6 different evaluators, resulting in 720 evaluations.

The annotators were presented with a translation-reference pair at a time. The two evaluation tasks corresponding to the same reference were presented at two different times with at least 40 other tasks in-between. This was done to prevent any possible spurious effects that may arise from remembering the content of a first translation, when evaluating the second translation of the same sentence. During each evaluation task, the evaluators were asked to assess the quality of a translation by providing a score between 0–100 (Graham et al., 2013). The observed inter-annotator agreement (Cohen's kappa) among our annotators was 0.321. This is slightly higher than the overall inter-annotator agreement of 0.284 reported in WMT'12 for the Spanish-English.[3] For reading patterns we use the EyeTribe eye-tracker at a sampling frequency of 30Hz. Please refer to Abdelali et al. (2016) for our Eye-Tracking setup and to know about *iAppraise*, an evaluation environment that supports eye-tracking.

3.1 Evaluation

In our evaluation, we used eye-tracking features to predict the quality of a translation in a pairwise scenario in a protocol similar to the one from WMT'12. First, we obtained the predicted scores $\hat{y}_A^k$, $\hat{y}_B^k$ for translations A and B when evaluated by evaluator k. Then, we computed the agreements w.r.t. the scores y_A^k, y_B^k provided by the evaluator for the same pair of translations. That is, we considered an agreement when rankings were in order, e.g. $\hat{y}_A^k > \hat{y}_B^k \iff y_A^k > y_B^k$. Otherwise, we considered it a disagreement. Finally, we computed Kendall's tau correlation coefficient as follows: $\tau = \frac{agg-dis}{agg+dis}$. We evaluated the performance using a 10-fold cross-validation. While the folds were selected randomly, we ensured that all translations corresponding to the same sentence were included in the same fold, to prevent any overlap between train and test.

4 Results

In this section, we first analyze the results of coherent feature sets to measure their predictive power and to validate the intuitions about the information they capture. Later, we use combination of features and assess their suitability as evaluation metrics.

4.1 Gaze as a translation quality predictor

In Table 1, we show the results for the predictive models trained on different feature sets. For simplicity, we divide the feature groups in: reference only features (I), translation only features (II), translation and reference features (III); and lexicalized features (IV). In the last group, we also add a tri-gram language model scores for comparison purposes.

Reference only features In section I of the table, we observe the prediction results for the models that only used features from the references. Unsurprisingly, most of these features lack the predictive power to determine whether translation A is better than translation B (τ from 0.06 to 0.13). One would expect that important phenomena that can be

[3] For a rough comparison only. Note that these two numbers are not *exactly* comparable given that they are calculated on different subsets of the same data. Still, there is a fair agreement between the our evaluators and the *expected wins* from WMT'12 (avg. pairwise kappa of 0.381)

observed only on the reference (e.g. the overall difficulty of the sentence), are neutralized in a pairwise setting, because an evaluator would read both instances of the reference text similarly.[4]
However, some features like the dwell time ($\tau = 0.13$) yield better results than others. This could be explained by the need to go back to the reference, when reading a confusing translation, thus spending more time reading the reference.

Translation only features In section II, we observe the results for the translation features. At a first glance, we realize that the correlation results are much higher than for the reference features (τ from 0.17 to 0.23). This supports the hypothesis that reading patterns can help to distinguish good from bad translations. Furthermore, it also supports specific intuitions about these reading patterns. For example, the fluency of a sentence is important (forward jumps, $\tau = 0.17$), but the number of regressions are better predictors of the quality of a sentence ($\tau = 0.22$). Additionally, the time spent reading a translation (dwell time) is a good predictor of the quality ($\tau = 0.22$). All of the above validate the intuition that reading patterns capture information about the quality of a translation. In general, using translation eye-tracking features in a pairwise evaluation, can help to predict which translation is better.

Translation and reference features Reference and translation features are not independent. Inter-region jumps capture the number of times that evaluators go between translation and references before making judgment. In section III, we observe that these features can be useful to predict the quality of a translation ($\tau = 0.18$).

Lexicalized features In the last rows of the table, we show that reading patterns help to evaluate more than just the fluency of a translation. A simple language model score (B_{LM}), is a weaker quality predictor ($\tau = 0.17$) than most of the eye-tracking translation features. Using the lexicalized version of the jump features gives additional predictive power ($\tau = 0.22$). Furthermore, by adding the total number of jumps and backward jumps to the LM features, we would obtain a considerable gain in correlation ($\tau = 0.30$). This suggests that the reading patterns capture information about more than just fluency.

SYS	Feature Sets (total features)	τ
I. Eye-tracking: Reference		
EyeRef$_{fj}$	Forward jumps (5)	0.06
EyeRef$_{bj}$	Backward jumps (5)	0.11
EyeRef$_{dist}$	Total jump distance (1)	0.09
EyeRef$_{visit}$	Total number of jumps (1)	0.10
EyeRef$_{time}$	Dwell time (1)	0.13
II. Eye-tracking: Translation		
EyeTra$_{fj}$	Forward jumps (5)	0.17
EyeTra$_{bj}$	Backward jumps (5)	0.22
EyeTra$_{dist}$	Total jump distance (1)	0.19
EyeTra$_{visit}$	Total number of jumps(1)	0.23
EyeTra$_{time}$	Dwell time (1)	0.22
III. Eye-tracking: Inter-region		
EyeInter	Jumps b/w regions (2)	0.18
IV. Lexicalized features		
B$_{LM}$	Language model (6)	0.17
EyeLex$_{all}$	Lexicalized gaze jumps combined (6)	0.22

Table 1: Results of individual eye-tracking features based on reference region, translation region, inter-region and lexicalized information

4.2 Gaze to build an evaluation metric

So far, we've shown that the individual sets of features based on reading patterns can help to predict translation quality, and that this goes beyond simple fluency. One question that remains to be answered is whether these features could be used as a whole to evaluate the quality of a translation *semi*-automatically. That is, whether we can use the gaze information, and other lexical information to anticipate the score that an evaluator will assign to a translation. Here, we present evaluation results combining several of these gaze features, and compare them against BLEU (Papineni et al., 2002), which uses lexical information and is designed to measure not only *fluency* but also *adequacy*.

In Table 2, we present results in the following way: in (I) we present the best non-lexicalized feature

[4] Although there could be differences based on corresponding translation, which may result in different values for the reference features.

combinations that improve the predictive power of the model. In (II) we re-introduce the results of lexicalized jumps feature. In (III) we present results of BLEU and the combination of eye-tracking features with it. Finally in (IV) we present the human-to-human agreement measured in average Kendall's tau and in max human-to-human Kendall's tau.

Combinations of translation jumps In section I we present several combinations of features. All of them include the backward jumps feature. This feature provides predictive power ($\tau = 0.22$), which is orthogonal to other features. This is in line with our initial hypothesis that for a bad translation, an evaluator needs to go back and forth several times to understand it. Combining the backward jumps with the total number of jumps (CTJ_1) slightly increases the correlation to $\tau = 0.25$. Adding the jump distance (CTJ_2) also increases its τ to 0.27. While this correlation is lower than BLEU ($\tau = 0.34$), it does showcase the predictive power of the reading patterns.

Combinations with BLEU When we combined BLEU with the translation jumps, we observed an increment in the τ to 0.37. Combining BLEU with the lexicalized jumps, yields the best combination ($\tau = 0.42$). Although moderate, these increments suggest that the reading patterns could be capturing additional phenomenon besides adequacy and fluency, such as structural complexity. These phenomena remain to be explored in future work.

Human performance On average, evaluators agreements with each other are fair ($\tau = 0.33$) and below the best combination (CB_3), while the maximum agreement of any two evaluators is relatively higher ($\tau = 0.53$). This tells us that on average the semi-automatic approach to evaluation that we propose here is already competitive to predictions done by another (average) human. However, there is still room for improvement with respect to the most-agreeing pair of evaluators.

5 Related Work

Eye-tracking devices have been used previously in the MT research. Stymne et al. (2012) used eye-tracking to identify and classify MT errors. Doherty et al. (2010) conducted a study using eye-tracking for MT evaluation and showed correlation between fixations and BLEU scores. Doherty and O'Brien (2014) evaluated the quality of machine translation output in terms of its *usability* by an end user. Guzmán et al. (2015) used eye-tracking to show that having monolingual environment improves the consistency of the evaluation.

SYS	**Feature Sets**	τ
I. Combination of translation jumps		
EyeTra_{bj}	Backward jumps	0.22
CTJ_1	Backward jumps, total jumps	0.25
CTJ_2	Backward jumps, total jumps, distance	0.27
II. Eye-tracking: Best Lexicalized		
EyeLex_{all}	Lexicalized gaze jumps	0.22
III. Combinations with BLEU		
B_{bleu}	BLEU	0.34
CB_1	B_{bleu} + EyeTra_{bj}	0.38
CB_2	B_{bleu} + CTJ_2	0.39
CB_3	B_{bleu} + EyeLex_{all}	0.42
IV. Human performance		
Avg	Avg. human-to-human agreement	0.33
Max	Max. human-to-human agreement	0.53

Table 2: Result of combining several jump and lexicalized features with BLEU. The column *Feature Sets* shows the name of the systems whose features are combined for that particular run. We also included the average and maximum observed *tau* between any two evaluators, as a reference.

Our work is different, as we: i) proposed novel eye-tracking features and ii) model gaze movements to predict human judgment.

6 Conclusion

We have shown that the reading patterns detected through eye-tracking can be used to predict human judgments of automatic translations. To this end, we extracted novel lexicalized and non-lexicalized features from the eye-tracking data motivated by notions of reading difficulty, and used them to predict the quality of a translation. We have shown that these features capture more than just the fluency of a translation, and provide complementary information to BLEU. In combination, these features can be used to produce semi-automatic metrics with improved the correlation with human judgments.

In the future, we plan to extend our experiments to a large set of users and different language pairs. Additionally we plan to improve the feature set to take into account phenomena such as *early termination*, i.e. when an evaluator makes a judgment before finishing reading a translation. We plan to deepen our analysis to determine what kind of information is being used beyond fluency and adequacy.

References

Ahmed Abdelali, Nadir Durrani, and Francisco Guzmán. 2016. iAppraise: A Manual Machine Translation Evaluation Environment Supporting Eye-tracking. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California.

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*, Prague, Czech Republic.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada. Association for Computational Linguistics.

Charles Clifton, Adrian Staub, and Keith Rayner. 2007. Eye Movements in Reading Words and Sentences. *Eye Movements: A Window on Mind and Brain*, pages 341–372.

Moreno I. Coco and Frank Keller. 2015. The Interaction of Visual and Linguistic Saliency during Syntactic Ambiguity Resolution. *The Quarterly Journal of Experimental Psychology*, 68(1):46–74.

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 Automatic MT Evaluation Technique. *Machine translation*, 24(1):1–13.

Nadir Durrani, Helmut Schmid, and Alexander Fraser. 2011. A Joint Sequence Translation Model with Integrated Reordering. In *Proceedings of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT'11)*, Portland, OR, USA.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22.

Simon Garrod. 2006. Psycholinguistic Research Methods. *The Encyclopedia of Language and Linguistics*, 2:251–257.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous Measurement Scales in Human Evaluation of Machine Translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, Sofia, Bulgaria.

Francisco Guzmán, Ahmed Abdelali, Irina Temnikova, Hassan Sajjad, and Stephan Vogel. 2015. How do Humans Evaluate Machine Translation. In *Proceedings of the 10th Workshop on Statistical Machine Translation*, Lisbon, Portugal.

Dan Witzner Hansen and Qiang Ji. 2010. In the Eye of the Beholder: A Survey of Models for Eyes and Gaze. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(3):478–500.

Trevor A Harley. 2013. *The Psychology of Language: From Data to Theory*. Psychology Press.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2001. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Marcel A. Just and Patricia A. Carpenter. 1980. A Theory of Reading: From Eye Fixations to Comprehension. *Psychological review*, 87(4):329.

Paul-Philipp Metzner. 2015. *Eye Movements and Brain Responses in Natural Reading*. Ph.D. thesis, University of Potsdam.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Association for Computational Linguistics*, ACL '02, pages 311–318, Philadelphia, PA, USA.

Mary C. Potter. 1983. Representational Buffers: The Eye-Mind Hypothesis in Picture Perception, Reading, and Visual Search. *Eye Movements in Reading: Perceptual and Language Processes*, pages 423–437.

Keith Rayner. 1998. Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological bulletin*, 124(3):372.

Elizabeth R. Schotter, Randy Tran, and Keith Rayner. 2014. Dont Believe What You Read (Only Once) Comprehension Is Supported by Regressions During Reading. *Psychological science*, page 0956797614531148.

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*, 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 *Proceedings of the International Conference on Language Resources and Evaluation*, Istanbul, Turkey.

Joseph Turian, Luke Shen, and I. Dan Melamed. 2003. Evaluation of Machine Translation and its Evaluation. In *Proceedings of Machine Translation Summit IX*, New Orleans, LA, USA.

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*, Columbia, Maryland, USA.