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

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute

جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

Hassan Sajjad, Francisco Guzman, Nadir Durrani, Houda Bouamor*, Ahmed Abdelali, Irina Temnikova, Stephan Vogel

Qatar Computing Research Institute, HBKU Qatar; Carnegie Mellon University, Qatar*

Introduction & Motivation

Problem: Human evaluation suffers from inter- and intra-annotator agreements

Evaluation scores are too subjective

Hypothesis: Reading patterns from evaluators can help to

- shed light into the evaluation process
- understand which parts of the sentences are difficult to evaluate
- develop a semi-automatic evaluation system based on reading patterns

Our Solution: use reading patterns as a method to distinguish between good and bad translations

In addition:

- identified novel features from gaze data
- model and predict the quality of translations as perceived by evaluators

Features

1. **Jump features (words transitions):** word-level forward and backward gaze jumps
2. **Total jump distance:** total gaze distance covered while evaluating
3. **Inter-region jumps:** gaze jumps between translation and the reference
4. **Dwell time:** longer time eyes spend on a region
5. **Lexicalized features:**
   - extract streams of lexical sequences R
   - score using a trigram language model

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

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

Model

- Linear regression model with ridge regularization
- Ridge coefficient $\hat{\beta}$ minimizes the error

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

- Parameter λ controls the amount of shrink applied to regression coefficients
- Used the glmnet package of R for cross-validation to find the best value of λ on the training data

Experimental Setup

Data

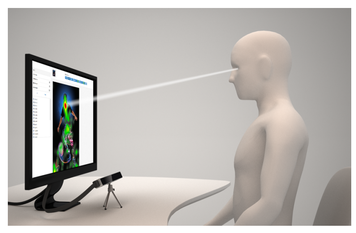
- Subset of the Spanish-English WMT'12 Evaluation task
- Selected 60 medium-length sentences, evaluated by at least 2 different annotators
- Selected the *best* & the *worst* translations, according to a human evaluation score, based on *expected wins*.
- Total 120 evaluation tasks x 6 different evaluators = 720 evaluations

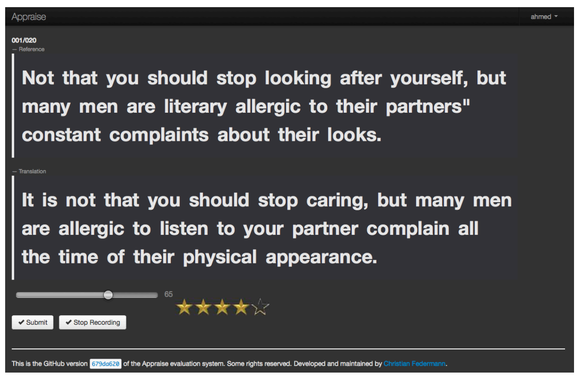
Eye-tracking Annotations

- Present evaluators with a translation-reference pair
- The best/worst translations of the same sentence have been shown with at least 40 different tasks in between
- Assign a 0-100 score to each task
- Inter-annotator kappa = 0.321 (slightly higher than the overall IAA in WMT'12 for Spanish-English – 0.284)

Tool

- EyeTribe eye-tracker
- Sampling frequency of 30Hz.
- Evaluation environment: *iAppraise*





Evaluation

Protocol similar to WMT'12

- Pairwise evaluation
- Computed the Kendall's tau coefficient
- Evaluated using 10-fold cross-validation

Results

Lack predictive power

Reading patterns on translation and inter-region bring useful information

Lexicalized gaze jumps brings additional value than an LM

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

Combining the best features with BLEU brings: reading patterns capture more than just fluency and adequacy

Combinations with BLEU	
B_{bleu}	0.34
B_{bleu} + EyeTra_{bj}	0.38
B_{bleu} + EyeLex_{all}	0.42

Conclusion

Conclusions: Eye–tracking features extracted captures additional information and can complement traditional measures (BLEU).

Future work: more users, language pairs, early termination features, deepen analysis.