Assessing the Impact of Translation Errors on MT Quality with Mixed-effects Models

Marcello Federico, Matteo Negri, Luisa Bentivogli, Marco Turchi
FBK – Fondazione Bruno Kessler, Trento, Italy
{federico, negri, bentivo, turchi}@fbk.eu

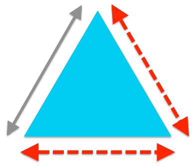
MOTIVATION

Support MT system development by analyzing the relations:

- between **MT errors and human quality judgments**
- between **MT errors and the sensitivity of automatic metrics**

...Most prior works focus on the relation (correlation) between **human judgments and automatic metrics**

Human judgments
Global quality assessments



Automatic metrics
Holistic view of system's behavior

Error annotations
Focused analysis of system's weaknesses

What error types have the highest impact on human quality judgments?

What error types have the highest impact on MT evaluation metrics?

What MT evaluation metrics show a sensitivity to errors more similar to humans?

MIXED LINEAR MODELS (MLMs)

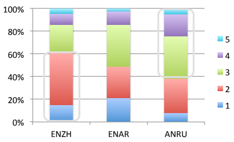
MLMs enhance conventional regression models by complementing *fixed effects* with *random effects* that absorb random variability inherent to the specific experimental setting that generates the observations (*i.e.* covariates that cannot be exhaustively observed)

DATA

- ~400 EN/ZH, EN/AR, EN/RU sentence pairs
- Translations produced by two anonymous MT systems
- Quality scores (1 to 5) assigned by three experts
- MT errors (*lex, morph, miss, reo*) annotated by one expert

VARIABILITY IN THE OBSERVATIONS

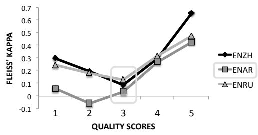
QUALITY SCORES



Distribution of quality scores

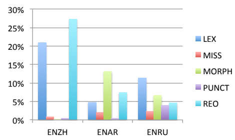
	Ann1	Ann2	Ann3	Sys1	Sys2
ENZH	2.38	2.69	2.21	2.29	2.56
ENAR	2.76	2.77	1.84	2.39	2.53
ENRU	2.82	2.72	2.96	2.87	2.79

Avg. score per annotator | system



Inter-annotator agreement

ERROR ANNOTATION



Distribution of error types

SYSTEM'S PERFORMANCE

	BLEU		TER		GTM	
	Sys1	Sys2	Sys1	Sys2	Sys1	Sys2
ENZH	27.95	44.11	64.52	48.13	62.15	72.30
ENAR	19.63	25.25	68.83	63.99	47.20	52.33
ENRU	27.10	31.07	60.89	54.41	53.74	56.41

Sentence-level automatic scores per system

ERRORS vs. QUALITY JUDGEMENTS

PREDICTION CAPABILITY

Task: predict human scores
Metric: MAE
MLMs compared to:

- 5 *univariate* models (baseline = sum of all error types)
- 2 *multivariate* models (all error types, with/without interactions)

Model	ENZH	ENAR	ENRU
baseline	0.58	0.73	0.67
lex	0.67	0.78	0.72
miss	0.72	0.89	0.74
morph	0.72	0.89	0.74
reo	0.70	0.82	0.76
FLM w/o Interact.	0.59	0.77	0.65
FLM	0.57	0.72	0.63
MLM	0.53	0.61	0.61

ERROR IMPACT

Slope coefficients as a measure of impact: highest decrement wrt intercept = highest impact)

Positive values for error combinations = combined impact is lower than the sum of the single errors

Model	ENZH	ENAR	ENRU
Intercept	4.29	3.79	4.21
lex	-1.27	**-0.96**	-1.12
miss	**-1.76**	-0.90	**-1.30**
morph	-0.48	-0.83	-0.51
reo	-1.01	-0.75	-0.18
lex:miss	1.00	0.39	0.68
lex:morph	-	0.29	0.32
lex:reo	0.50	0.21	-
miss:morph	-	0.35	-
miss:reo	0.54	0.33	-
morph:reo	-	0.37	-

ERRORS vs. AUTOMATIC METRICS

PREDICTION CAPABILITY

Task: predict BLEU, TER, GTM scores
Similar results: lowest MAE with MLMs

ERROR IMPACT

	BLEU			TER			GTM		
Error	ENZH	ENAR	ENRU	ENZH	ENAR	ENRU	ENZH	ENAR	ENRU
Intercept	60.55	38.45	51.73	32.41	52.25	33.40	83.57	60.11	75.38
lex	-18.78	-9.25	-16.57	16.87	9.66	18.45	-13.63	**-7.60**	-16.13
miss	-**23.20**	-10.41	-6.75	-	-	8.24	**-14.87**	-	-5.98
morph	-	-9.97	-12.65	-	8.90	11.41	-	-6.60	-10.42
reo	-13.27	-7.62	-10.57	14.44	9.81	6.39	-7.29	-5.50	-7.03
lex:miss	14.37	4.97	-	-	-	-	8.24	-	-
lex:morph	-	-	5.27	-	-	-5.22	-	-	4.92
lex:reo	8.57	3.57	5.40	-7.24	-4.35	-	5.46	3.22	3.65
miss:morph	-	4.44	-	-	-	-	-	-	-
miss:reo	6.74	-	4.30	-	-	-6.38	5.07	-	4.71
morph:reo	-	3.81	-	-	-4.97	-	-	2.57	-
Pearson	***0.98***	***0.97***	0.70	-0.58	-0.78	-0.78	***0.98***	0.78	0.74
Spearman	***0.97***	***0.91***	0.73	-0.57	-0.59	-0.80	***0.97***	0.59	0.76

The errors with highest impact vary across different translation directions

For some translation *directions*, some of the metrics show a sensitivity to errors similar to human judges

In some cases metrics and humans are most sensitive to the same error type

Error frequency does not correlate with human preferences (MLMs are more effective than methods based on raw error counts)

The impact of error interactions can be subject to measurable "discount" effects.
Sometimes with high correlation with humans, sometimes not

EMNLP 2014: Conference on Empirical Methods in Natural Language Processing — October 25–29, 2014 — Doha, Qatar