Research Group in Computational Linguistics

Translationese indicators for human translation quality estimation

(based on English-to-Russian translation of mass-media texts)

Maria Kunilovskaya

University of Wolverhampton

13 March, 2023

General description

'Shining-through' example

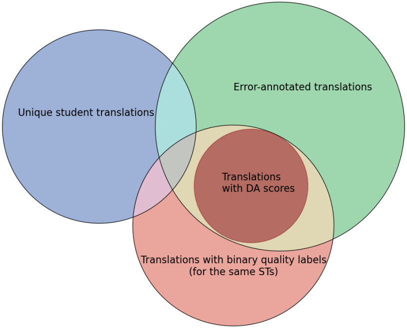
– How much time?
– Five hours.
– Such much?
– For whom how ...
– Finished injaz?
– Aaask!

Overarching aim: Explore the relations between translationese indicators and human translation quality.

Experimental setup: Human translation quality estimation task cast as text classification or regression problems, and feature analysis.

Research subcorpora

1. subsets from *Russian Learner Translator corpus* of various sizes by type of annotation



2. comparable professional translations: 404 parallel docs, 384 K words (BBC Russian Service, InoSMi, RNC);
3. comparable non-translations: 497 docs, 523 K words (RNC)

Quality labels/scores

Operational definitions of quality

- *Holistic judgments*: agreed assessment of competition jury/exam board in real life; top and bottom grades converted to 'bad', 'good' labels, verified in an additional annotation experiment ($\alpha = 0.524$, accuracy 91%).
- Scores from *error annotation* used as part of feedback to students in a real-life practical translation course, which implemented accuracy-fluency distinction (top-level category agreement: 80.5% of errors in the same location, $\alpha = 0.535$).
- *Direct assessment*: perceived quality for sentences presented in the context on a 100-point scale (documents: $\alpha = 0.541$, sentences: $\alpha = 0.463$)

+ Known *status* of translations produced by defined subjects (students, professionals).

Numeric representations

Proposed feature sets:

1. linguistically-motivated 60 morphosyntactic and textual features (from UD annotation)

Alternative representations:

- surface-based TF-IDF
- 4 types of sentence embeddings and word embeddings

+ (for quality-related experiments) MTQE features (QuEst++)

Methodology

Learning algorithms

- default linear Support Vector Machine
- one-layer neural network for quality control

Feature analysis

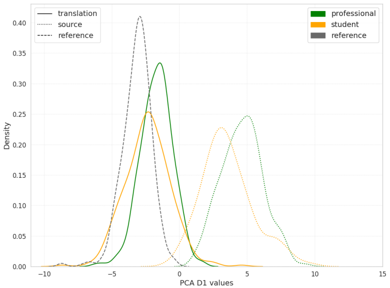
- recursive feature elimination
- univariate analysis (single-feature classifiers and regressors)
- statistical analyses
- PCA-based visualisations

Translationese indicators

How good are the UD features to capture translationese?

representation	Acc	F1
UD (prof)	90.34	90.22
mdeberta3 (prof)	98.44	98.36
UD (stud)	89.41	88.96
mdeberta3 (stud)	96.67	96.63

Translation detection task



What are the prominent translationese indicators?
(based on feature selection)

- longer and more complex sentences
- inflated frequencies of *additive discourse markers, analytical passives, copula verbs, modal predicates, personal pronouns, finite verbs, determiners*

Prominent trends and associated translation strategies
(based on statistical testing)

- shining through
- (over-)normalisation

Quality estimation tasks on quality labels (linear SVM)

professionals vs students

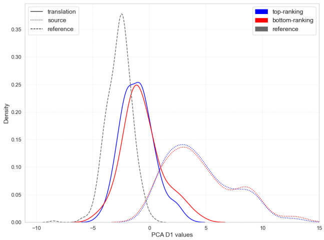
- dissimilar translationese patterns (UD F1=76.24)
- quality related distinctions (QuEst++ F1=83.00)
- topical differences overshadow translationese (tf-idf F1=89.59)

bad vs good

rep	Accuracy	F1
UD	61.39	61.00
quest61	47.22	46.96
ruRoberta	75.00	74.89
mdeberta3	78.33	78.14

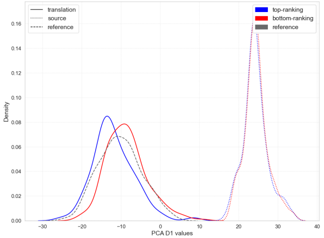
F1=68.9 on selected features

UD features

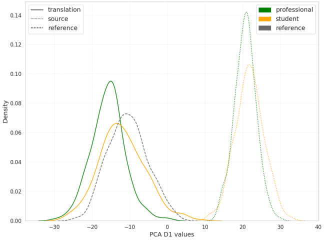


- SL/TL independent translationese features are important!
- Bad: longer sentences, complex sentence structure, lower TTR, analytical passives, more nouns as subjects, more modal predicates, more verbal (overuse of copula, deverbal nouns and participles)

mdeberta3 binary quality



mdeberta3: prof experience



Scores from error annotation (553 documents)

	accuracy		fluency		tq	
	r	RMSE	*r*	RMSE	*r*	RMSE
UD	**0.43**	0.95	**0.43**	1.18	**0.45**	1.72
quest61	0.37	0.98	0.42	1.16	0.36	1.73
tfidf	0.48	0.92	0.49	1.14	0.47	1.69
ruRoberta	0.51	0.91	0.53	1.08	0.54	1.57
mdeberta3	**0.58**	0.87	**0.58**	1.05	**0.62**	1.5

Regression results for *unweighted* error-based quality scores

Observations from feature analysis:

- no difference between accuracy and fluency (!)
- the very weak correlations between features and scores
- confusing observations for individual features

Scores from Direct Assessment (140 documents)

	da_mean	
rep	*r*	RMSE
UD	0.23	7.27
quest61	0.18	7.44
ruRoberta	0.22	7.35
mdeberta3	0.37	7.22

Results:

- none of the representations was more successful than the other in learning DA scores,
- in fair experimental setting, *mdeberta3* vectors performed better on some error-based scores than on DA scores,
- UD feature analysis is hardly reliable

Observations from sentence-level experiments

	Error-based scores		DA
	accuracy	fluency	da_mean
UD	0.17	0.23	0.29
quest70	0.14	0.25	0.33
ruRoberta	0.29	0.31	0.39
mdeberta3	0.27	0.33	0.39

Spearman's r on 3,224 sentence pairs (SVR)

- translationese-aware features were relatively competitive only for fluency scores (difference between accuracy and fluency!),
- an accidental finding:
  error-based quality scores reflected the properties of texts better than they fit the properties of sentences,
- interpretation of features does not make sense

Contributions

1. a theoretically-motivated feature set for translationese diagnostics in English-to-Russian mass-media translation;
2. evidence that lower-ranking translations exhibited more translationese than higher-ranking translations (UD features were competitive against other representations);
3. description of dissimilar translationese patterns in professional varieties;
4. evidence of dissimilarities between three quality assessment methods in terms of sensitivity to translationese and in terms of capturing document-level properties;
5. datasets for document- and sentence-level HTQE experiments in English-to-Russian language pair with three types of quality judgments

Theoretically disputable assumptions and limitations

1. translation quality does not boil down to fluency;
2. the approach is biased towards shining-through indicators with no distinction between negative and positive transfer;
3. the approach is heavily-dependent on register-comparability of translations and non-translations;
4. limited extent and reach of the study (findings apply to the given translation direction and register, given proposed features);
5. limited application of translationese approaches to sentence-level quality estimation

Thank you!

Translationese indicators
for human translation quality estimation

Maria Kunilovskaya