Linguistic features as a quality metric for human and automatic translation

Maria Kunilovskaya

Annual Progress Report 2019

RGCL, University of Wolverhampton

Novermber 07, 2019


x

x


Outline

Expected outcome of the PhD project

Linguistic features as a quality metric for human and automatic translation

Rank or classify multiple targets for the same source and/or produce an quality estimate for a single candidate, given the source and expected target language text fit (TL model)

My keywords:

- **area** (human) translation quality estimation (HTQE)
- **how** ML inc. NN, text vectorization
- **theory** translationese studies, variational linguistics, distributional semantics, text complexity and fluency
- **data** comparable and parallel corpora
- **challenges** OOM, low disk space

Primary data

"the beginning of any corpus study is the creation of the corpus itself ... The results are only as good as the corpus" (Sinclair, 1991)

Labeled quality: binary classes

- EN > RU
- 542 text pairs, 10K sentences, 200K tokens
- mass-media texts
- results of translation competitions, test, exams

Translational varieties

- RusLTC (EN>RU): 650 text pairs, 27K sents, 580K wds
- GroCO (EN>GE): 132 pairs, 15K sents, 320K wds
- argumentative and informative texts
- professional vs learner

Linguistic resources

>80% of the research time is spent on data collection and pre-processing

LM train corpora and reference

- Russian National Corpus (main and newspaper)
- EN and RU wikipedia (2019 dump)
- EN and RU comparable slices of web corpora (Aranea)

- symbol unification
- word- and sentence-tokenization
- stop words removal
- sentence length normalization
- lemmatization, lempos representation
- (bigram) NER

Research directions

Major research directions in 2019:

1. Building genre-comparable corpora
   - mono- and cross-lingual perspectives
   - evaluation: constructed resources and human judgment

   methods used: ML inc. neural networks, keyword analysis, clustering
2. Linguistic features and translationese
   - types of translationese in two target languages wrt competence levels
   - morpho-syntactic translationese as a quality measure

   methods used: UD-based feature engineering, PCA-LDA
3. Cross-linguistic text similarity and feature-less HTQE
   - accuracy module: semantic similarities for aligned texts on bilingual word vectors
   - fluency module: perplexities of rnn-based LM and ELMo last softmax layer

   methods: language modeling per se, models surprisal

Monolingual: get genre-comparable ST collections

Task: reconcile 8 balanced vs 10 unbalanced genres[1] in two corpora

CroCO and RusLTC 'essays' include

CroCo	Joint statement by M and N on renewables
RusLTC	BBC piece "Are work suits on the way out?"

Approach:

1. fit a (neural) model on genre-annotated data to produce text vectors that reflect text functions

   an essay, a speech, a pop-sci or an opinion

2. evaluate on 'known' genre-composition corpora
3. apply the best model to find functional clusters in the normative corpus and select the most similar texts to the centroid of the targeted cluster

▸ Do you want to know more abt the resources used for training and evaluation?

(Kunilovskaya and Sharoff, 2019)

[1] ▸ genre variation guesser

Monolingual: Results

Best model: a multi-task biLSTMa learner which back-propagates the accumulated loss for 10 tasks

Best intrinsic evaluation results on multi-hot transforms of the target and the model predictions (10-folds cv, macro F1 score):

EN 0.841
RU 0.849
(gained from stacking mixed vectors and Biber's features[2])

Classification on 'known' corpora:
(better than Biber's or keywords and more informative)

EN 0.79
RU 0.70
(stacking mixed vectors and Biber's features adds 0.02)

Resulting conundrum: genre-comparable or big data?

[2] NB! comparable Biber's features extraction for RU is not a trivial matter

Cross-lingual: build TL reference corpus

Task: What is your expected TL text fit, given the ST?

Evaluation Results: are the predictions returned by independent models directly comparable? Test on known similarity parallel and comparable texts!

* based on Euclidean distance as similarity measure

expected similarity

	category	similarity	mean
set 1	fiction	.432	.470
	media	.476	
	ted	.456	
	pop-sci	.514	
set 2	fiction	.315	.305
	media	.263	
	ted	.323	
	pop-sci	.317	
set 3	academic	.396	.214
	fiction	.259	
	non-academic	.127	
	personal	.139	
	promotion	.145	
	reportage	.216	
set 4	academic::fict	-.190	.004
	non-ac::promo	.116	
	pers::report	.085	

measured similarity

Linguistic features: interpretable translationese detection

'Translationese' feature set

- features shared by EN, DE, RU
- motivated by previous CBTS research and translationese studies
- limited to UD-based morpho-syntax
- use the best-suited treebanks (2.1, 2.2, 2.3)

Applied to the two research tasks:

- Compare two language pairs and two translational varieties on the amount and type of translationese captured by the suggested feature set (Kunilovskaya, Lapshinova (2019) Translation in Transition4 talk)
- Predict hand-annotated translation quality (Kunilovskaya and Lapshinova-Koltunski, 2019)

Final feature set (42 items)

8 morph. forms:

comp, sup, shortpassive, bypassive, infs, pverbals, deverbals, finites

7 morph. categories:

ppron, demdets, possdet, indef, mquantif, cconj, sconj

7 UD relations:

acl, aux, aux:pass, ccomp, nsubj:pass, parataxis, xcomp

3 synt. functions:

attrib, copula, nnargs

8 sentence type and structure:

simple, numcls, neg, relativ, pied, correl, mpred, whconj

2 graph-based:

mhd, mdd

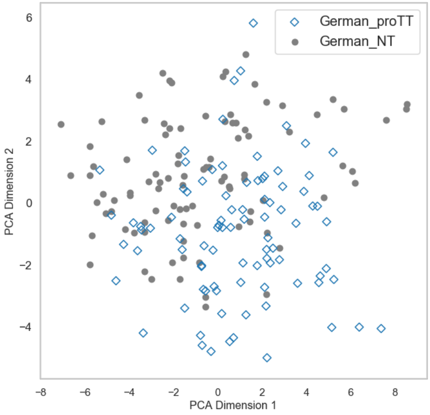
6 sem. types of discourse markers:

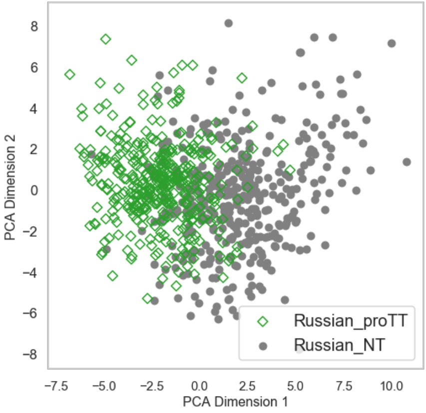
addit, advers, caus, tempseq, epist and but

1 measure of lexical density:

lexTTR

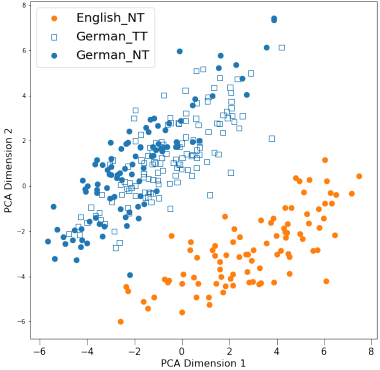
Do our features capture translationese?

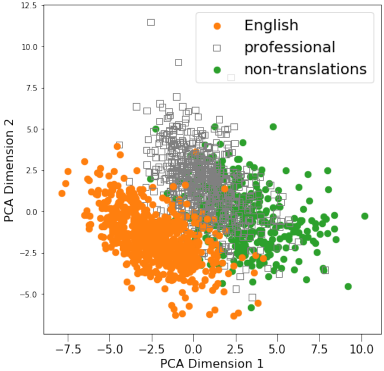




SVM classification for (balanced) translations and non-translations:
EN-DE pair: F1 = 0.79 EN-RU pair: F1 = 0.91

How do translations fit in the SL/TL gap?





Note:

(1) the shining-through shift towards the SL and

(2) the upward shift of translations.

Ask me how useful the features were for revealing

▸ types of translationese in DE and RU

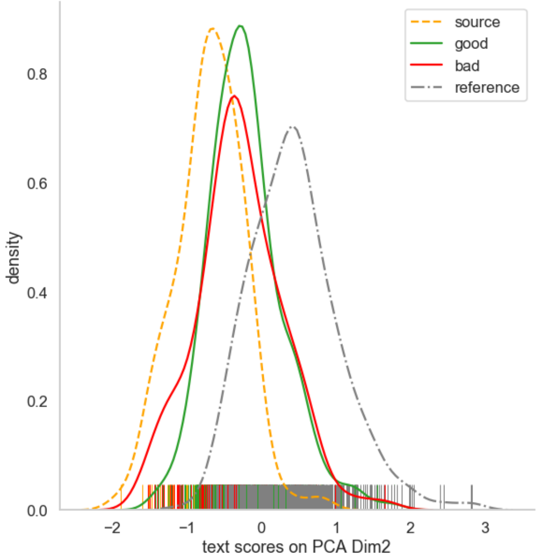
Student / professional vs. non-translations in Russian

	precision	recall	f1-score
pro	0.91	0.94	0.93
ref	0.94	0.91	0.92
macro avg	0.92	0.92	0.92
stu	0.93	0.95	0.94
ref	0.94	0.92	0.93
macro avg	0.94	0.94	0.94

Best indicators of translationese

`possdet, whconj, relativ, correl, lexdens, lexTTR, finites, deverbals, sconj, but, comp, numcls, simple, nnargs, ccomp`

Good and bad translations vs. non-translations



Translationese features as quality indicators

	precision	recall	f1-score
bad	0.48	0.55	0.51
good	0.79	0.74	0.76
macro avg	0.63	0.64	0.64

FYI: stratified dummy accuracy 0.52; macro-F1 0.49

Most informative for bad vs. good distinction

`copula, finites, pasttense, infs, relativ, lexdens, addit, ccomp, but, sconj, nnargs, acl, advers, ppron, sentlength`

The intersection with the 15 top translationese indicators includes: `finites, lexdens, but, relativ, nnargs, sconj, ccomp`

Word vectors instead of linguistic features

The distributional approach to model language use is too tempting to be ignored.

Accuracy module

- Embeddings capture the (distributional) properties of words
- Representations in two languages can be transformed into a shared semantic space
- Task: find a way to calculate semantic similarity for aligned texts that reflects accuracy

Fluency module

- Language models (LM) calculate the probability of a vocabulary item to be next in a sequence
- Bad (disfluent) translations have higher entropy and should result in higher LM perplexity
- Use LM perplexity as a fluency measure

Ask me about the results involving word vectors

Other ideas to predict quality

- Smooth Inverse Frequency to get text vectors (Arora et al., 2017; Ranashinghe et al., 2019)
- Quest++ features
- lexical features (nltk.collocations, gensim Phrase, Phraser)
- sentence-level predictions?
- error-annotation informed approaches?

Thank you very much

PhD thesis provisional title:

Linguistic features as a quality metric
for human and automatic translation

Questions?

Maria Kunilovskaya

References I

Arora, S., Liang, Y., and Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In ICLR 2017.

Kunilovskaya, M. and Lapshinova-Koltunski, E. (2019). Translationese Features as Indicators of Quality in English-Russian Human Translation. In Proceedings ofthe 2nd Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019), pages 47–56.

Kunilovskaya, M. and Sharoff, S. (2019). Towards Functionally Similar Corpus Resources for Translation. In Proceedings of Recent Advances in Natural Language Processing, pages 583–592.

References II

Kutuzov, A. and Kuzmenko, E. (2017). WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models, pages 155–161. Springer International Publishing, Cham.

Ranashinghe, T., Orasan, C., and Mitkov, R. (2019). Enhancing unsupervised sentence similarity methods with deep contextualised word representations. In Proceedings of Recent Advances in Natural Language Processing, pages 994–1003. RANLP.

Sharoff, S. (2018). Functional Text Dimensions for annotation of Web corpora. Corpora, 13(1):65–95.

Sinclair, J. (1991). Corpus, Concordance, Collocation. Oxford University Press.

Appendix0. Genres: give it a try!

Match sentences and typical genre labels:

texts	labels
Seventy four primary school teachers participated in a cross-sectional survey conducted in Western Australia. Teachers' attitudes and efficacy toward integration of students with disabilities were measured using the Opinions Relative to Integration of Students with Disabilities scale.	argument
America's stubborn retention of the death penalty is usually seen as the abolitionist movement's greatest defeat. And yet in the long term it may prove to be one of its greatest assets. If even America, with its complex legal guarantees...	academic
"Miss Peregrine's etc., etc." isn't just an explosion at the imagination well. It's a Deepwater Horizon of ideas, a flaming wreck of a rig that spews so much creativity in every direction I was ducking for cover, scrambling for an escape hatch.	encyclopedia
Economies of scale are factors that cause the average cost of producing something to fall as the volume of its output increases. Hence it might cost 3,000 to produce 100 copies of a magazine but only 4,000 to produce 1,000 copies.	evaluation
And 17 years later I did go to college. But I naively chose a college that was almost as expensive as Stanford, and all of my working-class parents' savings were being spent on my college tuition. After six months, I couldn't see the value in it.	legal
These terms and conditions operate to the exclusion of any terms and conditions put forward by the customer. No variations to these terms and conditions shall be binding unless agreed in writing..	personal

go back

Monolingual: Resources used

Training text representation model

hand-labeled data from Functional Text Dimensions (Sharoff, 2018) (task: how much the text resembles each of the 10 suggested prototypes)

ID	argument	fiction	instruction	news	legal	...
text12	2	0	0.5	1	0	...

Krippendorff's alpha >0.76 English/Russian: 1,624/1,930 texts, >2 M tokens each

Evaluation corpora

Six function-motivated categories (aligned with 6 FTDs) from

- BNC (David Lee's scheme)
- Russian National Corpus (RNC)

Appendix1: Types of translationese: analysis and indicators

Statistical univariate analysis

provided that a feature is indicative of translationese in general

- shining-thru features = in the gap btw the SL and TL
  ex. deverbals
  EN(0.149)—PRO(0.183)—STU(0.252)—REF(0.336)
- SL/TL-independent features = away from SL and TL
  ex. subordinate conjunctions
  REF(0.284)—EN(0.45)—STU(0.515)—PRO(0.576)
  esp. the features for which there is no language gap

Multivariate analysis

shining-through indicators as the intersection between 20 best language contrast indicators and 20 best translationese indicators

Resulting feature sets: Four types of features

EN > DE (pro)

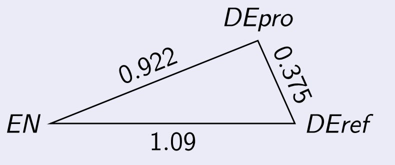
- 10 fully adapted in translation (SL≠TLref, but TLref=TLpro)
- 4 features useless for this analysis (SL=TLref and TLref=TLpro)
- 19 shining-thru indicators (features in the language gap)
- 8 SL/TL-independent translationese indicators (features distinct from both languages and outside of the gap between them)

EN > RU (pro)

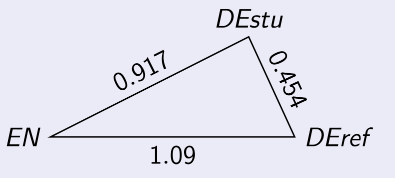
- 2 fully adapted in translation (SL≠TLref, but TLref=TLpro)
- 2 features useless for this analysis (SL=TLref and TLref=TLpro)
- 21 shining-thru indicators (features in the language gap)
- 16 SL/TL-independent translationese indicators (features distinct from both languages and outside of the gap between them)

Translationese measure: Euclidean distances

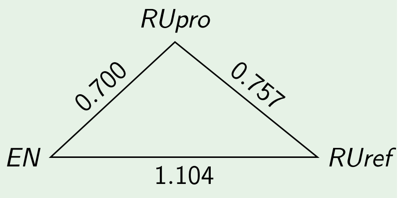
professional



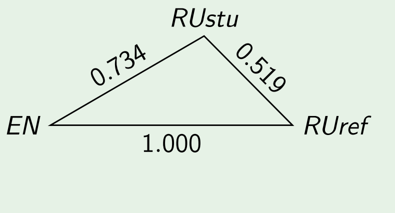
student



professional



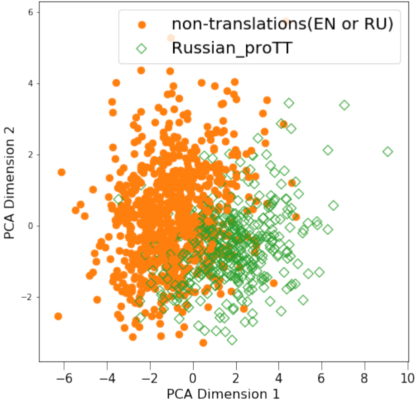
student



Note the horizontal and the vertical shifts of the translations corner!

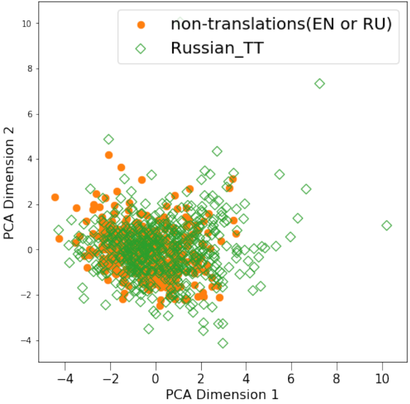
EN-RU idiosyncrasy for SL/TL-independent translationese

PROFESSIONALS



SVM-classifying (EN+TLref) vs pro: F1=0.815

STUDENTS



SVM-classifying (EN+TLref) vs stu: F1=0.591

Results on translational norms, varieties and types of translationese

1. professional norms: more translationese in EN>RU than in EN>DE (based on distances to the TL reference);
2. feature sets for shining-through and SL/TL-independent translationese effects for each pair (ask me how we distilled ▸ indicators);
3. 67-91% feature lists overlap for the varieties in both language pairs;
4. weird: in EN>RU students data is devoid of the specificity found in professional texts;
5. in EN>DE professionalism is less translationese, but we don't have enough data to make rigorous statistic inferences

Building and evaluating resources

Task: build own linguistic resources (useful for fluency module, too)

- genre-comparable vectors for fluency
- lempos WITH stopwords, inc. learn bilingual transform

Materials to preprocess, annotate and learn vectors from:

- EN wiki2019 dump (26 GiB .gz, 122M sents, 2.4G tokens)
- RU ruscorpora+wiki2018 dump (55M sents, 867M tokens)
- RNC newspaper subcorpus (12M sents, 202M tokens)

Evaluation

- monolingual vectors
  Spearman's ρ on enSimLex-999/ruSimLex-965

	en	ru
CommonCrawl tokens	.371	.308
wiki lempos	.401	.321
wiki lempos func	.413	.311
wiki shared space	.413	.311
newspaper5	—	.315

- bilingual vectors
  on bilingual glossary (1.5K word pairs), cosine, model=100K

	P@1	P@5	@10
CommonCrawl tokens			
wiki lempos	65%	79%	83%
~~wiki lempos func~~	~~0.3%~~	~~1.3%~~	~~2.1%~~

Cross-linguistic text similarity as accuracy measure

Initial and improved baseline (vectors: CommonCrawl tokens; UD lempos, no stopwords(punct)/as is)

FYI: stratified dummy accuracy 0.52; macro-F1 0.49

		representation	**learner**	**macro F1**
	1	tf-idf scaled BOW	SVM[3], gridsearch (C:100, gamma:1, features:1000, norm:'l1', idf:True)	char(3,3): .674; word(3,3):.652 pos(3,3): .377
	2	45 translationese	SVM UD-based feats	.642
word vectors	3	siamese ST&TT with dot product as Dense layer input (tried stacking, Manhattan, Eucledian)	biLSTM, (units=128; bilingual vectors, 0.1 val_split, patience=5, computed class weights, batch_size=1)	lempos:.630/ .599
	4	cosine for ST&TT	SVM	lempos:.579/ .526
	5	summed ST&TT	SVM; bilingual vectors	lempos: .607/.608

[3] 'balanced', kernel='rbf', stratified 10-fold cv

Language Models perplexity as fluency feature

Is (probability -> entropy -> perplexity) indicative of quality?

how: perplexities from cross-entropies

$$H(p, q) = -\sum_{x \in \mathcal{X}} p(x) \log q(x) \qquad (1)$$

RNN-based trigram LM

- learnt on genre-comparable corpus (for toy-model trained on a corpus of 1697 texts, lempos, inc. functionals/punct)
- macro-F1: 0.56

toy results only

embeddings from LM (ELMo)

- ELMo for what it is in essence: an LM at the last layer
- an ELMo pre-trained on wiki+ruscorpora(989M, tokens) (Kutuzov and Kuzmenko, 2017)
- results on toy 24 bad/good: .50 (best of the 4 classifier)!