Translation Task Difficulty and Target Text Properties

Maria Kunilovskaya, E. Lapshinova-Koltunski, E. Teich

Keywords: `translationese, subtrees, translation task difficulty, information theory (IT)`

1. Why this research?

Translationese studies do not directly account for **cross-lingual** nature of translation. → Factor in source text (ST) comprehension and cross-lingual transfer difficulty!

Previous attempts using average sentence surprisal for translation detection failed. → Use **subsentencial** units!

2. Hypothesis and Expectations

🤔 Is the translation task difficulty linked to the linguistic specificity of translations?

👉 If yes, translationese is a rational response to the increased cognitive pressure, while producing deviant translations requires less effort [1].

3. Definitions and Reserach design

IT-based translation difficulty for operational units

Operational units: Aligned content words and their subtrees.

ST comprehension effort
- unit surprisal from GPT2

ST-TT transfer effort
- unit translation entropy
- unit align/similarity score

Each ST-TT pair gets a vector of indices, averaged across constituent operational units.

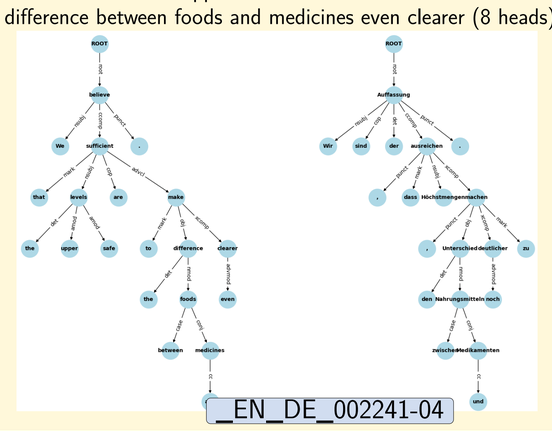
IT indices for subtrees and words

A subtree is a NVAA head with its dependents of depth 1.
- cosine-align (SBert+[2])
- translation entropy for:

Wir sind der Auffassung ausreichen
PRON sein DET Auffassung ccomp

NVAA content words:
- alignment: AWESoME [3]
- surprisal: GPT2 [4, 5]

We believe that the upper safe levels are sufficient to make the difference between foods and medicines even clearer (8 heads)



Alternative ST comprehension effort: syntax

- mean dependency distance
- mean hierarchical distance
- tree depth
- branching factor

Measure of translationese

Response variable: Probability of being a translation on a strong SVM translationese classifier (on delexicalised features) [6].

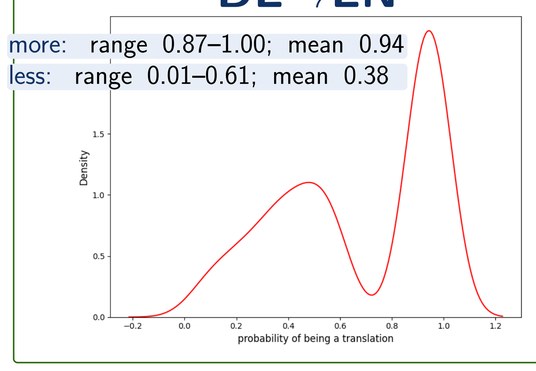
DE⇔EN data: Europarl-UdS(>450): 1.5Kdocs, ≈40Ksegs.

F1 score: 88.79% DE; 79.72% EN

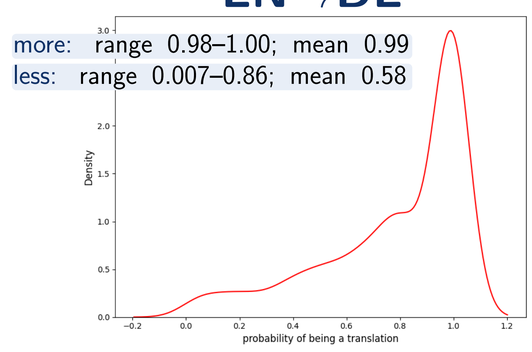
Also tried distilbert-base-cased: aks me about the outcome :-)

Distribution of probabilities in 1000 top and bottom targets

DE→EN



EN→DE



Experimental setup

SVM with a linear kernel
Feature selection with RFE
10-fold cross-validation
Data shuffling and scaling
+ distilbert on ST (baseline)

Also tried:
Neural encoder+regressor on unit sequences as 3D vectors [surprisal, align, entropy].
Result: no discernible trend.

4. Regression Results

	approach	unit	Pearson	MAE
deen	distilbert	NA	0.21	0.29
ende			0.20	0.23
deen	**syntax**	NA	0.28±0.10	0.24±0.01
ende			0.21±0.06	0.21±0.01
deen	IT	trees	0.13±0.08	0.27±0.02
ende			-0.09±0.05	0.22±0.02
deen	IT	**words**	0.23±0.10	0.25±0.01
ende			0.13±0.13	0.21±0.02
deen	IT	trees+words	0.23±0.09	0.20±0.01
ende			0.06±0.08	0.21±0.02
deen	IT+syntax(best)	words	0.31±0.09	0.19±0.01
ende			0.21±0.06	0.16±0.01

Bonus: Spoken data (165/137 document pairs)

	approach	unit	Pearson	MAE
deen	IT+syntax(best)	words	0.23±0.26	0.16±0.04
ende			0.39±0.21	0.16±0.04

6. Summary and Takeaways

💡 Regardless of approach, the association between task difficulty and translationese is weak: Limited confirmation for the hypothesis.

💡 translation entropy and align scores are more related to translationese than ST surprisal.

💡 ST syntactic complexity measures perform better than IT-based and neural features.

💡 The link between task difficulty and translationese is stronger in translated English.

💡 Trees as operational units did not live up to expectations.

💡 IT indices for content words only perform better.

7. Acknowledgments & References

This research is funded by German Research Foundation (DFG), Project-ID 232722074 – SFB 1102.

[1] Michael Carl and Moritz Jonas Schaeffer. Why translation is diffcult: A corpus-based study of non-literality in post-editing and from-scratch translation. *Hermes (Denmark)*, (56):43–57, 2017.

[2] Nils Reimers and Iryna Gurevych. paraphrase-multilingual-minilm-l12-v2: Multilingual sentence embeddings. Pretrained model on `huggingface.co`, 2020. Available at `https://www.sbert.net`.

[3] Zi-Yi Dou and Graham Neubig. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, 2021.

[4] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI*, 2019.

[5] Stefan Schweter. German gpt-2 model. *Zenodo*, November 2020.

[6] Maria Kunilovskaya, Koel Dutta Chowdhury, Heike Przybyl, Cristina España i Bonet, and Josef Van Genabith. Mitigating Translationese with GPT-4: Strategies and Performance. In *Proceedings of the 25th Annual conference of the European Association for Machine Translation*, pages 411–430, Sheffield, UK, 24–27 June 2024. Association for Computational Linguistics.