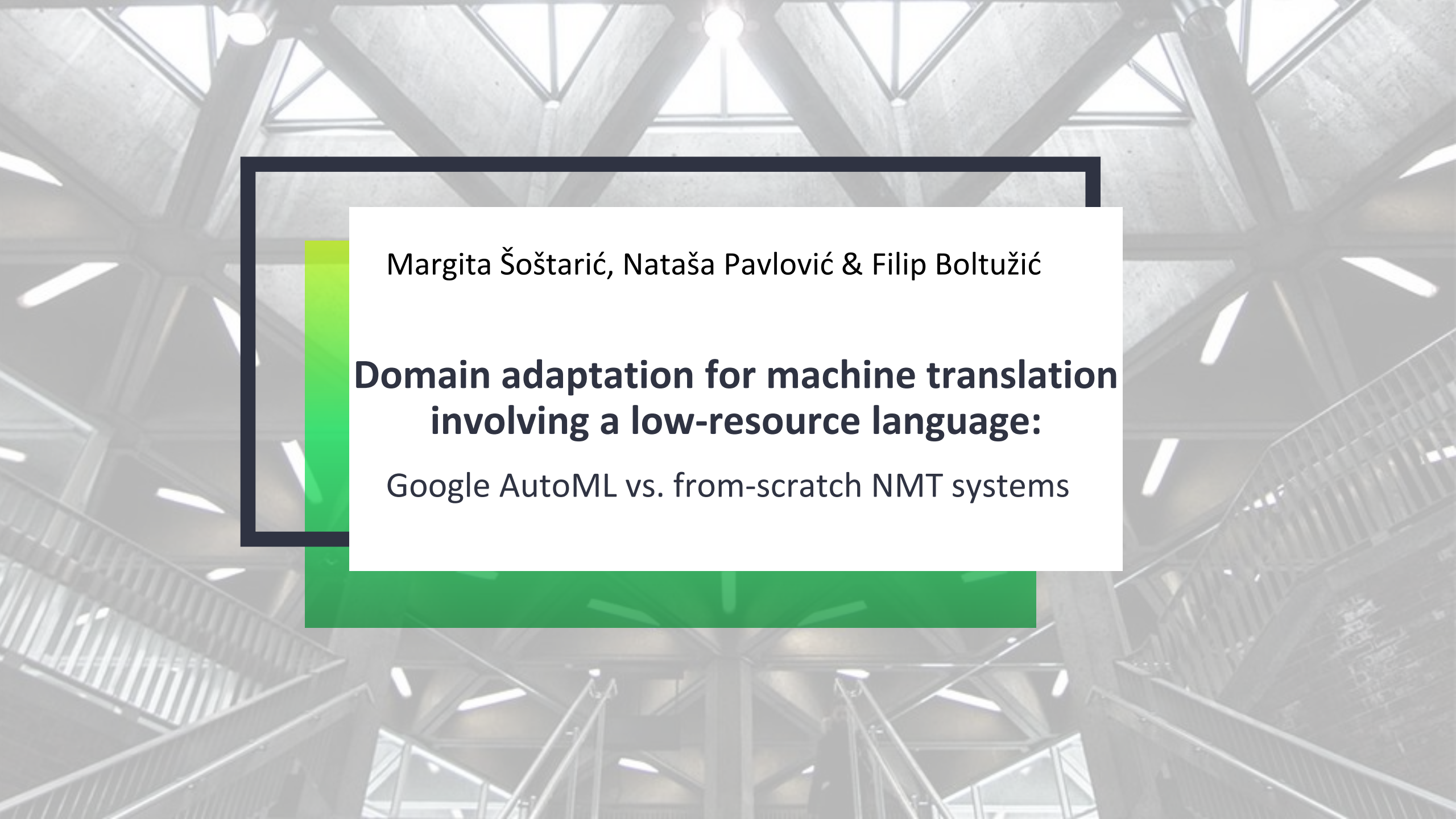


Margita Šoštarić, Nataša Pavlović & Filip Boltužić

Domain adaptation for machine translation involving a low-resource language:

Google AutoML vs. from-scratch NMT systems

OVERVIEW

- Motivation for the study
- Approach and experimental setup
- Automatic and human evaluation
- Discussion
- Conclusion

PROBLEM

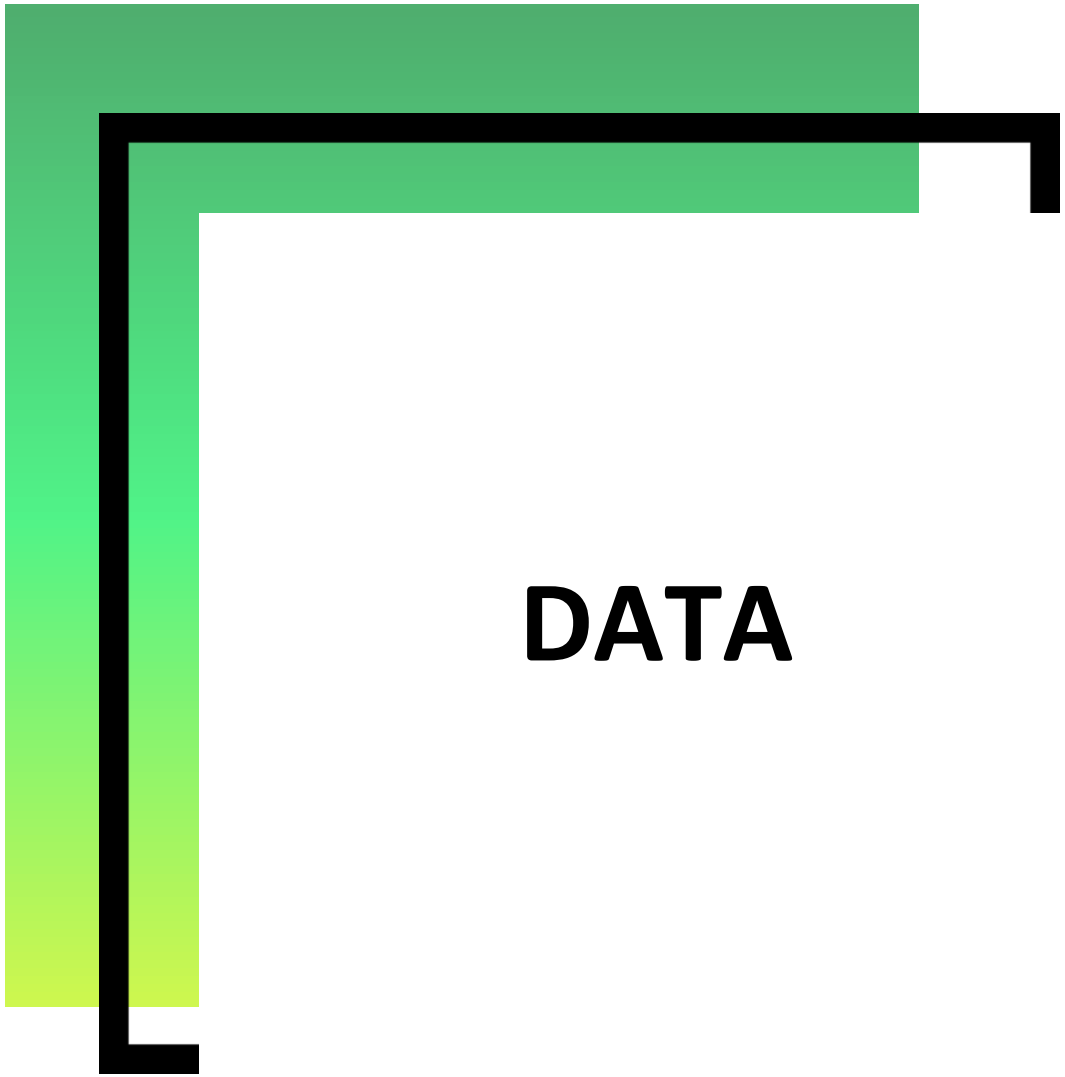
- Adaptation of MT systems for use in specific domains
- Acute for low-resource languages (e.g. Croatian)
- Relevant for small translation companies and freelancers (resources, costs, expertise, time… present barriers)

POSSIBLE SOLUTIONS

- Build own MT models from scratch
  - general-purpose model
  - model adapted to a specific domain
- Use commercial but affordable models
  - off the shelf (e.g. Google Translate)
  - adapted to a specific domain (e.g. using Google Cloud AutoML)

AIM OF THE STUDY

- To assess these possible solutions in terms of:
  - translation quality
  - accessibility to small translation companies and freelancers


DATA

DATASETS USED IN THE STUDY

- In domain
  - EN-HR medical corpus from Cochrane (high quality)
  - EN-HR medical glossary (12K entries)
- Out of domain
  - EN-HR parallel corpora available on Opus corpus, ELRC-SHARE, CLARIN; JRC-Acquis
- Set-up easily reproducible in translation companies

DATA PRE-PROCESSING

- Standard data preparation and filtering methods applied
  - 1,715,169 unique segment pairs of out-of-domain data
  - 45,674 unique segment pairs of in-domain data

	Train	Validation	Test
Out of domain	1,707,169	4,000	4,000
In domain	42,674	1,000	2,000

- Segmenting into sub-word units: byte-pair encoding (Sennrich et al., 2016b)

SYSTEM TRAINING

FROM-SCRATCH MODELS

- Transformer architecture (Vaswani et al., 2017)
  - self-attention mechanism
- OpenNMT toolkit (Klein et al., 2018)
  - used recommended training parameters
- Popular domain-adaptation methods:
  - fine-tuning (Luong & Manning, 2015; Sennrich et al., 2016a; Freitag & Al-Onaizan, 2016)
  - domain tagging (Kobus et al., 2017; Chu et al., 2017)
  - creating synthetic corpora (Sennrich et al., 2016a)

FROM-SCRATCH MODELS

TF_{gen} - general-purpose model with all available data, in and out of domain

TF_{med} - domain-adapted model, partly based on Chu et al. (2017)

- using in-domain and out-of-domain data - the ratio of learning from in-domain data increased by *oversampling*
- used medical glossaries to identify sentence pairs with EN terms on the source side and their HR equivalents on the target side, adding them to the in-domain data in oversampling
- training took 1.5 days

GOOGLE-BASED MODELS

GT_{gen} - general-purpose Google Translate

GT_{med} - domain-adapted model based on Google Translate

- adaptation using Google Cloud AutoML
- only requirement: uploading the in-domain dataset
- commercial, but affordable ($76.00/hour of training)
- training took 3 hours

- Can be integrated in a CAT tool with an API at low cost

EVALUATION

AUTOMATIC EVALUATION

- Used automatic metrics:
  - BLEU (Papineni et al., 2002)
  - chrF3 (Popović, 2015)
  - TER (Snover et al., 2006)

- Evaluated performance on out-of-domain and in-domain data
- Pre-processing steps reverted

The results of automatic evaluation

	Out of domain			In domain		
	BLEU	chrF3	TER	BLEU	chrF3	TER
GT_{gen}	28.63*	55.40	0.604*	22.40*	52.08*	0.619*
GT_{med}	22.13*	49.98*	0.673*	24.00	53.86	0.604
TF_{gen}	33.17	55.26	0.568	25.08	53.83	0.592
TF_{med}	**33.60**	**55.96**	**0.565**	**27.96***	**55.58***	**0.577***

The best results are bolded. Only the scores that are statistically significant ($p < 0.05$) in comparison to all other systems are marked with asterisks.

HUMAN EVALUATION

- Evaluators:
  - 27 in total: translation students, professional translators, medical professionals, medical students; all had experience in medical translation
- Evaluation tasks: source sentence + two machine translations
- Three options:
  - 1st translation is better
  - 2nd translation is better
  - both are equally good/bad
- Two criteria: accuracy and usability

WHY ACCURACY AND USABILITY?

Source sentence	There is little doubt that women should be encouraged to utilise positions which give them the greatest comfort, control and benefit during first stage **labour**.
GT_{gen} translation	Nema sumnje da žene treba poticati da koriste položaje koji im pružaju najveću udobnost, kontrolu i dobrobit tijekom prve faze **rada**. [There is no doubt that women should be encouraged to utilise positions which give them the greatest comfort, control, and benefit during the first stage of **work**.]

HUMAN EVALUATION

- Customized surveys for each evaluator
  - individual and shared subset, both with 10 source sentences and their respective translations
  - all 4 translations of the source sentence compared by the same evaluator
  - evaluations for the translations of 285 sentences
- Full system ranking determined from the averaged scores of Elo rating
  - calculated for each group of questions relating to the same source sentence

RESULTS OF HUMAN EVALUATION

	Accuracy		Usability	
	Elo rating	Rank	Elo rating	Rank
GT_{gen}	1201.65	1	1202.09	1
GT_{med}	1200.08	3	1199.68	3
TF_{gen}	1197.91	4	1198.19	4
TF_{med}	1200.34	2	1200.02	2

Inter-rater agreement (Fleiss' kappa): "fair" for both accuracy (0.301) and usability (0.237)

DISCUSSION

DISCUSSION OF RESULTS

- Discrepancy in system ranking for in-domain data:
  - automatic evaluation: TF_{med} – TF_{gen} / GT_{med} – GT_{gen}
  - human evaluation (both criteria): GT_{gen} – TF_{med} – GT_{med} – TF_{gen}
- Variation in evaluators' responses
  - closer examination of the data: quality varies within and between translations produced by all systems

OTHER OBSERVATIONS

- One third of evaluators ranked both GT systems above TF systems on both accuracy and usability
- Ranking for the common questions: $TF_{med} – GT_{gen} – GT_{med} – TF_{gen}$
- The performance of GT_{med} significantly dropped on out-of-domain data

SUMMARY

- Explored several simple methods to obtain general-purpose and domain-adapted models
- Using a service: Google Cloud AutoML
  - only requires basic computer literacy
  - paid service
- Training own systems: OpenNMT
  - the training and data pre- and post-processing take longer
  - basic programming skills and special resources required

CONCLUSIONS

- General-purpose and domain-adapted models performed similarly
  - automatic evaluation favoured from-scratch systems
  - human feedback varied
- All of the compared approaches are viable options for small translation companies and freelancers
- The issue of domain-adapted MT still far from resolved

THANK YOU FOR YOUR ATTENTION!

Margita Šoštarić
Cerence Inc.
Aachen, Germany
gita.sostaric@gmail.com

Nataša Pavlović
University of Zagreb
Zagreb, Croatia
npavlovi@ffzg.hr

Filip Boltužić
University of Zagreb
Zagreb, Croatia
filip.boltuzic@fer.hr

REFERENCES

Chu Chenhui, Raj Dabre, and Sadao Kurohashi. 2017. An Empirical Comparison of Simple Domain Adaptation Methods for Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Volume 2: Short Papers, pages 385–391.

Freitag, Markus, and Yaser Al-Onaizan. 2016. Fast Domain Adaptation for Neural Machine Translation. https://arxiv.org/pdf/1612.06897.pdf

Klein, Guillaume, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander M. Rush. 2018. OpenNMT: Neural Machine Translation Toolkit. In P*roceedings of the 13th Conference of the Association for Machine Translation in the Americas*, pages 177–184.

Kobus, Catherine, Josep Maria Crego, and Jean Senellart. 2017. Domain Control for Neural Machine Translation. In *Proceedings of Recent Advances in Natural Language Processing*, pages 372–378.

Luong, Minh-Thang, and Christopher D. Manning. 2015. Stanford Neural Machine Translation Systems for Spoken Language Domains. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 76-79.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311-318.

REFERENCES

Popović, Maja. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392-395.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016a. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 86-96.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016b. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223-231.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, pages 6000-6010.

TRAINING PARAMETERS

- number of layers: 6, number of heads: 8
- word vector size: 512, vocabulary size: 80000
- position encoding
- batch size: 4096, batch type: tokens
- optimization: adam, label smoothing: 0.1
- train steps: 200000, warm-up steps: 8000, learning rate: 2
- GPU: 2 (2080 RTx Ti)
- validation and model saving: every 10000 steps
  - last four saved models ensembled for translation
- cost: cloud-hosted $109 per model, GPU purchase $1200