NATIONAL RESEARCH
UNIVERSITY

School of Foreign Languages

MACHINE TRANSLATION METHODS FOR CREATING A CORPUS OF SIMPLIFIED TEXTS

Project Proposal

Research Advisor: Ekaterina Artemova, Associate Professor

Author: Aleksandra Izhevskaia
BFL1711

PLAN OF THE PRESENTATION

- Research Background
- Literature review
- Purpose + research gap
- Methods
- Expected outcomes

Time limit: 6-7 min

RESEARCH BACKGROUND

Sentence simplification:

They are culturally akin to the coastal peoples of Papua New Guinea.

↓

Their culture is similar to the culture of the coastal peoples of Papua New Guinea.

Key points about sentence simplification

- Can be beneficial for people with cognitive disabilities and language learners
- Can be done automatically with special tools
- Needs parallel corpora
- May benefit from machine translation

LITERATURE REVIEW

Relevant studies

- Nishihara, Kajiwara & Arase (2019). Controllable text simplification with lexical constraint loss
- Martin, Fan, de La Clergerie, Bordes & Sagot (2020). Multilingual unsupervised sentence simplification
- Coster,& Kauchak (2011). Simple English Wikipedia: a new text simplification task

Investigated issues:

- Encoder-decoder approach to the task
- Task transferring (using translation models for simplification)
- Availability of parallel corpora in English

Limitations:

- Language and task transferring for many languages, including Russian
- Creation of a high-quality parallel corpus in many languages, including Russian

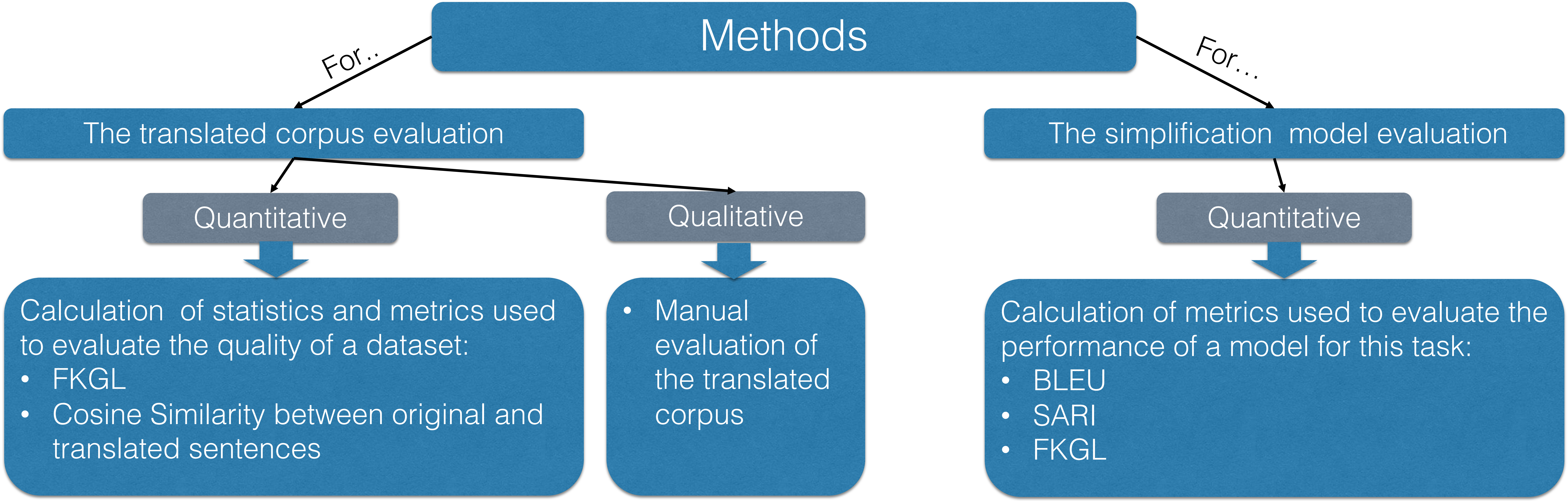
PURPOSE + RESEARCH GAP

To investigate the machine translation role in parallel corpora translation and facilitating sentence simplification for the Russian language

To address the gap

- To better understand machine translation in general and in the context of sentence simplification
- To provide a parallel simplification corpus translated to Russian

METHODS



Methods

For..

The translated corpus evaluation

Quantitative

Calculation of statistics and metrics used to evaluate the quality of a dataset:

- FKGL
- Cosine Similarity between original and translated sentences

Qualitative

- Manual evaluation of the translated corpus

For…

The simplification model evaluation

Quantitative

Calculation of metrics used to evaluate the performance of a model for this task:

- BLEU
- SARI
- FKGL

EXPECTED OUTCOMES

- Train a simplification model on both the original and the translated data and achieve a high quality of sentence simplification

- Show that the automatically translated data could be used successfully for training simplification models

- Identify common drawbacks of automatically translated data and aspects that may need additional manual correction

CONCLUSION

- Usefulness for the translation of foreign corpora and creation of original Russian parallel simplification corpora

- Contribution to the improvement of cross-lingual models for machine translation and sentence simplification

REFERENCES:

Andreeva M., Solnyshkina M., Solovyev V., Zaikin A., & Bukach O. (2020). Computing descriptive metrics and propositions in reading texts and recalls. In CEUR Workshop Proceedings.

Bahdanau, D., Cho, K. H., & Bengio, Y. (2015, September 1). Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. https://arxiv.org/abs/1409.0473v7

Brunato, D., Cimino, A., Dell'Orletta, F., & Venturi, G. (2016, November). Paccss-it: A parallel corpus of complex-simple sentences for automatic text simplification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (pp. 351-361).

Coster, W., & Kauchak, D. (2011, June). Simple English Wikipedia: a new text simplification task. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (pp. 665-669). Flesch, R. (1948). A new readability yardstick. Journal of Applied Psychology, 32(3), 221–233. https://doi.org/10.1037/h0057532

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2020). Language-agnostic bert sentence embedding. arXiv preprint arXiv:2007.01852.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A. F., Bogoychev, N., Martins, A. F. T., & Birch, A. (2018). Marian: Fast Neural Machine Translation in C++. ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of System Demonstrations, 116–121. http://arxiv.org/abs/1804.00344

Katsuta, A., & Yamamoto, K. (2018, May). Crowdsourced corpus of sentence simplification with core vocabulary. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Kuratov, Y., & Arkhipov, M. (2019). Adaptation of deep bidirectional multilingual transformers for Russian language. arXiv preprint arXiv:1905.07213.

Li, J., Lester, C., Zhao, X., Ding, Y., Jiang, Y., & Vydiswaran, V. G. V. (2020). PharmMT: A Neural Machine Translation Approach to Simplify Prescription Directions. Findings of the Association for Computational Linguistics: EMNLP 2020, 2785–2796. https://doi.org/10.18653/v1/2020.findings-emnlp.251

REFERENCES:

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., & Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics, 8, 726-742.

Martin, L., Fan, A., de La Clergerie, E., Bordes, A., & Sagot, B. (2020). Multilingual unsupervised sentence simplification. In arXiv. arXiv. https://commoncrawl.org

Nishihara, D., Kajiwara, T., & Arase, Y. (2019, July). Controllable text simplification with lexical constraint loss. In Proceedings of the 57th annual meeting of the association for computational linguistics: Student research workshop (pp. 260-266).

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. ArXiv, 21, 1–67. http://arxiv.org/abs/1910.10683

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 2017-December, 5999–6009. https://arxiv.org/abs/1706.03762v5

Wu, Y., Schuster, M., Chen, Z., Le, Q. v., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., … Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. http://arxiv.org/abs/1609.08144

Xu, W., Callison-Burch, C., & Napoles, C. (2015). Problems in Current Text Simplification Research: New Data Can Help. Transactions of the Association for Computational Linguistics, 3, 283–297.

Xu, W., Napoles, C., Pavlick, E., Chen, Q., & Callison-Burch, C. (2016). Optimizing Statistical Machine Translation for Text Simplification. Transactions of the Association for Computational Linguistics, 4, 401–415.

Zhang, X., & Lapata, M. (2017). Sentence Simplification with Deep Reinforcement Learning. EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings, 584–594. http://arxiv.org/abs/1703.10931

Thank you for your attention!

NATIONAL RESEARCH UNIVERSITY

School of Foreign Languages

MACHINE TRANSLATION METHODS FOR CREATING A CORPUS OF SIMPLIFIED TEXTS

Project Proposal

Research Advisor: Ekaterina Artemova, Associate Professor

Author: Aleksandra Izhevskaia
BFL1711