FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

ABSTRACT OF DOCTORAL THESIS

Zdeněk Kasner

Data-to-Text Generation with Neural Language Models

Institute of Formal and Applied Linguistics

Supervisor: Mgr. et Mgr. Ondřej Dušek, Ph.D.

Study Program: Computational Linguistics

Prague 2024

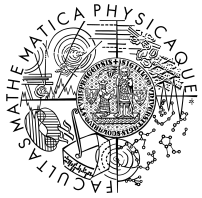
The results of this thesis were achieved in the period of a doctoral study at the Faculty of Mathematics and Physics, Charles University in years 2019–2024.

Doctoral Candidate:	Ing. Zdeněk Kasner Institute of Formal and Applied Linguistics
Supervisor:	Mgr. et Mgr. Ondřej Dušek, Ph.D. Institute of Formal and Applied Linguistics
Department:	Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University Malostranské náměstí 25, 118 00 Prague 1, Czech Republic
Opponents:	prof. dr. Emiel Krahmer School of Humanities and Digital Sciences Department of Communication and Cognition Tilburg University Tilburg, Netherlands dr. Yaji Sripada School of Natural and Computing Sciences University of Aberdeen Aberdeen, United Kingdom
Chairman of Academic Council:	doc. RNDr. Pavel Pecina, Ph.D Institute of Formal and Applied Linguistics

This abstract was distributed on July 24, 2024.

The thesis defense will take place on September 5, 2024 at 9:30 a.m. in front of a committee for thesis defenses in the branch Mathematical Linguistics at the Faculty of Mathematics and Physics, Charles University, Malostranské nám. 25, Prague 1, room S1.

The thesis can be viewed at the Study Department of Doctoral Studies of the Faculty of Mathematics and Physics, Charles University, Ke Karlovu 3, Prague 2.



MATEMATICKO-FYZIKÁLNÍ
FAKULTA
Univerzita Karlova

AUTOREFERÁT DISERTAČNÍ PRÁCE

Zdeněk Kasner

Generování textu z dat s neuronovými jazykovými modely

Ústav formální a aplikované lingvistiky

Školitel: Mgr. et Mgr. Ondřej Dušek, Ph.D.

Studijní program: Matematická lingvistika

Praha 2024

Disertační práce byla vypracována na základě výsledků získaných během doktorského studia na Matematicko-fyzikální fakultě Univerzity Karlovy v letech 2019–2024.

Doktorand:	Ing. Zdeněk Kasner Ústav formální a aplikované lingvistiky
Školitel:	Mgr. et Mgr. Ondřej Dušek, Ph.D. Ústav formální a aplikované lingvistiky
Školicí pracoviště:	Ústav formální a aplikované lingvistiky, Matematicko-fyzikální fakulta, Univerzita Karlova Malostranské náměstí 25, 118 00 Praha 1, Česká republika
Oponenti:	prof. dr. Emiel Krahmer School of Humanities and Digital Sciences Department of Communication and Cognition Tilburg University Tilburg, Nizozemsko
	dr. Yaji Sripada School of Natural and Computing Sciences University of Aberdeen Aberdeen, Velká Británie
Předseda RDSO:	doc. RNDr. Pavel Pecina, Ph.D Ústav formální a aplikované lingvistiky

Autoreferát byl rozeslán dne 24. července 2024.

Obhajoba disertační práce se koná dne 5. září 2024 v 09:30 před komisí pro obhajoby disertačních prací v oboru Matematická lingvistika na Matematicko-fyzikální fakultě UK, Malostranské nám. 25, Praha 1, v místnosti S1.

S disertační prací je možno se seznámit na studijním oddělení Matematicko-fyzikální fakulty UK, Ke Karlovu 3, Praha 2.

(Kasner and Dušek, 2022)

Bibliography

Dušek, O. – Kasner, Z. Evaluating Semantic Accuracy of Data-to-Text Generation with Natural Language Inference. In *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020*, p. 131–137, Dublin, Ireland, 2020. doi: 10.18653/V1/2020.INLG-1.19. Available at: https://doi.org/10.18653/v1/2020.inlg-1.19.

Kasner, Z. – Dušek, O. Neural Pipeline for Zero-Shot Data-to-Text Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022*, p. 3914–3932, Dublin, Ireland, 2022. doi: 10.18653/V1/2022.ACL-LONG.271. Available at: https://doi.org/10.18653/v1/2022.acl-long.271.

Kasner, Z. – Dušek, O. Data-to-Text Generation with Iterative Text Editing. In *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020*, p. 60–67, Dublin, Ireland, 2020a. doi: 10.18653/V1/2020.INLG-1.9. Available at: https://doi.org/10.18653/v1/2020.inlg-1.9.

Kasner, Z. – Dušek, O. Beyond Traditional Benchmarks: Analyzing Behaviors of Open LLMs on Data-to-Text Generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024. Available at: http://arxiv.org/abs/2401.10186. To appear.

Kasner, Z. – Dušek, O. Train Hard, Finetune Easy: Multilingual Denoising for RDF-to-Text Generation. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, p. 171–176, Dublin, Ireland (Virtual), 12 2020b. Available at: https://aclanthology.org/2020.webnlg-1.20.

Kasner, Z. – Mille, S. – Dušek, O. Text-in-Context: Token-Level Error Detection for Table-to-Text Generation. In *Proceedings of the 14th International Conference on Natural Language Generation, INLG 2021*, p. 259–265, Aberdeen, Scotland, UK, 2021. doi: 10.18653/V1/2021.INLG-1.25. Available at: https://doi.org/10.18653/v1/2021.inlg-1.25.

Kasner, Z. – Garanina, E. – Plátek, O. – Dušek, O. TabGenie: A Toolkit for Table-to-Text Generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2023*, p. 444–455, Toronto, Canada, 2023a. doi: 10.18653/V1/2023.ACL-DEMO.42. Available at: https://doi.org/10.18653/v1/2023.acl-demo.42.

Kasner, Z. – Konstas, I. – Dušek, O. Mind the Labels: Describing Relations in Knowledge Graphs With Pretrained Models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik*, p. 2390–2407, Croatia, 2023b. doi: 10.18653/V1/2023.EACL-MAIN.176. Available at: https://doi.org/10.18653/v1/2023.eacl-main.176.

List of Publications

Kasner, Z. – Dušek, O. Train Hard, Finetune Easy: Multilingual Denoising for RDF-to-Text Generation. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, p. 171–176, Dublin, Ireland (Virtual), 12 2020b. Available at: https://aclanthology.org/2020.webnlg-1.20

- Citations (without self-citations): 9

Kasner, Z. – Dušek, O. Data-to-Text Generation with Iterative Text Editing. In *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020*, p. 60–67, Dublin, Ireland, 2020a. doi: 10.18653/V1/2020.INLG-1.9. Available at: https://doi.org/10.18653/v1/2020.inlg-1.9

- Citations (without self-citations): 17

Kasner, Z. – Dušek, O. Neural Pipeline for Zero-Shot Data-to-Text Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022*, p. 3914–3932, Dublin, Ireland, 2022. doi: 10.18653/V1/2022.ACL-LONG.271. Available at: https://doi.org/10.18653/v1/2022.acl-long.271

- Citations (without self-citations): 21

Dušek, O. – Kasner, Z. Evaluating Semantic Accuracy of Data-to-Text Generation with Natural Language Inference. In *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020*, p. 131–137, Dublin, Ireland, 2020. doi: 10.18653/V1/2020.INLG-1.19. Available at: https://doi.org/10.18653/v1/2020.inlg-1.19

- Citations (without self-citations): 47

Kasner, Z. – Mille, S. – Dušek, O. Text-in-Context: Token-Level Error Detection for Table-to-Text Generation. In *Proceedings of the 14th International Conference on Natural Language Generation, INLG 2021*, p. 259–265, Aberdeen, Scotland, UK, 2021. doi: 10.18653/V1/2021.INLG-1.25. Available at: https://doi.org/10.18653/v1/2021.inlg-1.25

- Citations (without self-citations): 6

Kasner, Z. – Garanina, E. – Plátek, O. – Dušek, O. TabGenie: A Toolkit for Table-to-Text Generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2023*, p. 444–455, Toronto, Canada, 2023a. doi: 10.18653/V1/2023.ACL-DEMO.42. Available at: https://doi.org/10.18653/v1/2023.acl-demo.42

- Citations (without self-citations): 2

Kasner, Z. – Konstas, I. – Dušek, O. Mind the Labels: Describing Relations in Knowledge Graphs With Pretrained Models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik*, p. 2390–2407, Croatia, 2023b. doi: 10.18653/V1/2023.EACL-MAIN.176. Available at: https://doi.org/10.18653/v1/2023.eacl-main.176

- The analysis of verbalizing relations in knowledge graphs with pretrained language models (**??**).
- Citations (without self-citations): 3

Kasner, Z. – Dušek, O. Beyond Traditional Benchmarks: Analyzing Behaviors of Open LLMs on Data-to-Text Generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024. Available at: http://arxiv.org/abs/2401.10186. To appear

- The analysis of data-to-text generation with open large language models (**??**).
- Citations (without self-citations): 2

Only publications relevant to this thesis are included. The number of citations was computed using Semantic Scholar API. Total number of citations of publications related to the topic of the thesis (without self-citations) by June 14, 2024: **107**.