FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

DOCTORAL THESIS

Jindřich Helcl

On the Importance of Context in Neural Machine Translation

Institute of Formal and Applied Linguistics

Supervisor: prof. RNDr. Jan Hajič, Dr.

Study Program: Computer Science
Specialization: Computational Linguistics

Prague 2019

I declare that I carried out this doctoral thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

Prague, March 21, 2019 Jindřich Helcl

Title: On the Importance of Context in Neural Machine Translation

Author: Jindřich Helcl

Department: Institute of Formal and Applied Linguistics

Supervisor: prof. RNDr. Jan Hajič, Dr.,
Institute of Formal and Applied Linguistics

Abstract:

anstract abstract abstract

Keywords: multimodal machine translation, neural machnie translation, combining language and vision, deep learning

Název práce: Význam kontextu v neuronovém strojovém překladu

Autor: Jindřich Helcl

Katedra: Ústav formální a aplikované lingvistiky

Vedoucí práce: prof. RNDr. Jan Hajič, Dr.,
Ústav formální a aplikované lingvistiky

Abstrakt:

abstrakt abstrakt abstrakt

Klíčová slova: neautoregresivní strojový překlad, multimodální strojový překlad, neuronový strojový překlad, hluboké učení

Acknowledgements

děkuju. tos řek hezky.

TODO akcknowledgementy

Contents

1 Introduction

Ahoj, hello, world. Natural Language Processing (NLP)

Bibliography

Helcl, J. – Libovický, J. CUNI System for the WMT17 Multimodal Translation Task. In *Proceedings of the Second Conference on Machine Translation (WMT), Volume 2: Shared Task Papers*, p. 450–457, Copenhagen, Denmark, September 2017a. Association for Computational Linguistics.

Helcl, J. – Libovický, J. Neural Monkey: An Open-source Tool for Sequence Learning. *The Prague Bulletin of Mathematical Linguistics*. Apr 2017b, 107, 1, p. 5–17. ISSN 0032-6585.

Helcl, J. – Libovický, J. – Kocmi, T. – Musil, T. – Cífka, O. – Variš, D. – Bojar, O. Neural Monkey: The Current State and Beyond. In *Proceedings of the 13th Conference of The Association for Machine Translation in the Americas, Vol. 1: MT Researchers' Track*, p. 168–176, Boston, MA, USA, March 2018a. The Association for Machine Translation in the Americas.

Helcl, J. – Libovický, J. – Variš, D. CUNI System for the WMT18 Multimodal Translation Task. In *Proceedings of the Third Conference on Machine Translation*, p. 622–629, Brussels, Belgium, October 2018b. Association for Computational Linguistics.

Libovický, J. – Helcl, J. Attention Strategies for Multi-Source Sequence-to-Sequence Learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, p. 196–202, Vancouver, Canada, July 2017. Association for Computational Linguistics.

Libovický, J. – Helcl, J. – Tlustý, M. – Bojar, O. – Pecina, P. CUNI System for WMT16 Automatic Post-Editing and Multimodal Translation Tasks. In *Proceedings of the First Conference on Machine Translation*, p. 646–654, Berlin, Germany, August 2016. Association for Computational Linguistics.

Libovický, J. – Helcl, J. – Mareček, D. Input Combination Strategies for Multi-Source Transformer Decoder. In *Proceedings of the Third Conference on Machine Translation*, p. 253–260, Brussels, Belgium, October 2018. Association for Computational Linguistics.

List of Publications

TODO tohle se musí přepsat

Libovický, J. – Helcl, J. – Tlustý, M. – Bojar, O. – Pecina, P. CUNI System for WMT16 Automatic Post-Editing and Multimodal Translation Tasks. In *Proceedings of the First Conference on Machine Translation*, p. 646–654, Berlin, Germany, August 2016. Association for Computational Linguistics

- The paper describes a submission to the WMT16 which describes our early experiments with Multimodal Machine Translation (MMT). This paper is partially discussed in Sections **??** and **??**.
- Citations (without self-citations): 28

Libovický, J. – Helcl, J. Attention Strategies for Multi-Source Sequence-to-Sequence Learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, p. 196–202, Vancouver, Canada, July 2017. Association for Computational Linguistics

- The paper introduces techniques for combining multiple different inputs in sequence-to-sequence learning using Recurrent Neural Networks (RNNs). Content of this paper is discussed mainly in Section **??**.
- Awarded as an Outstanding Paper at ACL 2017 and ÚFAL Best Paper 2017.
- Citations (without self-citations): 19

Helcl, J. – Libovický, J. Neural Monkey: An Open-source Tool for Sequence Learning. *The Prague Bulletin of Mathematical Linguistics*. Apr 2017b, 107, 1, p. 5–17. ISSN 0032-6585

- This paper introduces a software tool *Neural Monkey* which was used for all the experiments in this thesis.
- Citations (without self-citations): 22

Helcl, J. – Libovický, J. CUNI System for the WMT17 Multimodal Translation Task. In *Proceedings of the Second Conference on Machine Translation (WMT), Volume 2: Shared Task Papers*, p. 450–457, Copenhagen, Denmark, September 2017a. Association for Computational Linguistics

- A submission to WMT17 MMT task. The submission tests architectures proposed in the previous paper in a more competitive setup and discusses techniques for acquiring additional training data which are discussed in Section **??**.
- Citations (without self-citations): 7

Helcl, J. – Libovický, J. – Kocmi, T. – Musil, T. – Cífka, O. – Variš, D. – Bojar, O. Neural Monkey: The Current State and Beyond. In *Proceedings of the 13th Conference of The Association for Machine Translation in the Americas, Vol. 1: MT Researchers' Track*, p. 168–176, Boston, MA, USA, March 2018a. The Association for Machine Translation in the Americas

- A paper summarizing the Neural Monkey software at the beginning of 2018.
- Citations (without self-citations): 1

Libovický, J. – Helcl, J. – Mareček, D. Input Combination Strategies for Multi-Source Transformer Decoder. In *Proceedings of the Third Conference on Machine Translation*, p. 253–260, Brussels, Belgium, October 2018. Association for Computational Linguistics

- The paper introduces techniques for input combinations in sequence-to-sequence models with self-attentive encoder and decoder.
- Citations (without self-citations): 0

Helcl, J. – Libovický, J. – Variš, D. CUNI System for the WMT18 Multimodal Translation Task. In *Proceedings of the Third Conference on Machine Translation*, p. 622–629, Brussels, Belgium, October 2018b. Association for Computational Linguistics

- A paper summarizing the Neural Monkey software at the beginning of 2018.
- Citations (without self-citations): 0

Only publication relevant to this thesis are included. The number of citations was computed using Google Scholar. Total number of citations of publication related to the topic of the thesis (without self-citations): 77 (by the thesis submission on March 21, 2019).

List of Abbreviations

MMT Multimodal Machine Translation. 5, 6

NLP Natural Language Processing. 1

RNN Recurrent Neural Network. 5

List of Tables

List of Figures