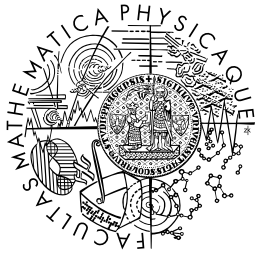

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Bohdan Ihnatchenko

Multi-Target Machine Translation

Institute of Formal and Applied Linguistics

Supervisor of the master thesis: doc. RNDr. Bojar Ondřej, Ph.D.
Study programme: Computer Science
Study branch: Artificial Intelligence

Prague 2020

This is not a part of the electronic version of the thesis, do not scan!

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

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

In ..... date ....

Author's signature

Dedication.

Title: Multi-Target Machine Translation

Author: Bohdan Ihnatchenko

Institute: Institute of Formal and Applied Linguistics

Supervisor: doc. RNDr. Bojar Ondřej, Ph.D., Institute of Formal and Applied Linguistics

Abstract: Abstract.

Keywords: Machine translation words

Contents

Introduction **2**

1 Background **3**
1.1 History of machine translation . . . 3
1.2 Transformer model . . . 3
1.3 Translation evaluation . . . 3

2 Experiment setup **4**
2.1 Dataset . . . 4
2.2 Training . . . 4
2.3 Experiments . . . 4

3 Random choise of target languages **5**
3.1 Performance drop on massively multilingual setup . . . 5
3.2 Performance decrease on richer data sets . . . 5

4 Group by language groups **6**
4.1 Language groups . . . 6
4.1.1 Germanic group . . . 6
4.1.2 Slavic with cyrillic script . . . 6
4.2 Selecting target languages by linguistic similarity . . . 6

Conclusion **7**

Bibliography **8**

List of Figures **9**

List of Tables **10**

List of Abbreviations **11**

A Attachments **12**
A.1 First Attachment . . . 12

Introduction

1. Background

1.1 History of machine translation

1.2 Transformer model

1.3 Translation evaluation

2. Experiment setup

2.1 Dataset

2.2 Training

2.3 Experiments

3. Random choise of target languages

3.1 Performance drop on massively multilingual setup

1-to-3, 5, 7, etc. models on en-to-36 dataset (0.9 mil. sentences per target language)

3.2 Performance decrease on richer data sets

1 to 3, 4, 5 on UN corpus (much more sentence pairs per target language)

4. Group by language groups

4.1 Language groups

1 to 2, 3, 4, 5, etc. models on en-to-36 dataset (0.9 mil. sentences per target language) compared with random runs

4.1.1 Germanic group

4.1.2 Slavic with cyrillic script

4.2 Selecting target languages by linguistic similarity

Conclusion

Bibliography

List of Figures

List of Tables

List of Abbreviations

A. Attachments

A.1 First Attachment