

Contents

FRONT MATTER	i
List of Figures	ix
List of Tables	xi
Abstract	xiii
Zusammenfassung	xv
Acknowledgements	xvii
CHAPTER 1	
Introduction	1
1.1 Motivation & Problem	1
1.2 Research Gap	6
1.3 Research Objective	7
1.4 Thesis Outline	8
1.4.1 Publications	8
1.4.2 Research Path	9
CHAPTER 2	
Mathematical Information Retrieval	17
2.1 Background and Overview	18
2.2 Mathematical Formats and Their Conversions	19
2.2.1 Web Formats	20
2.2.2 Word Processor Formats	25
2.2.3 Computable Formats	32
2.2.4 Images and Tree Representations	34
2.2.5 Math Embeddings	37
2.3 From Presentation to Content Languages	38
2.3.1 Background	39
2.3.2 Benchmarking MathML	43
2.3.3 Evaluation of Context-Agnostic Conversion Tools	48
2.3.4 Summary of MathML Converters	51
2.4 Mathematical Information Retrieval for LaTeX Translations	51

CHAPTER 3

Semantification of Mathematical LaTeX

3.1

Semantification via Math-Word Embeddings

57

3.1.1

Foundations and Related Work

61

3.1.2

Semantic Knowledge Extraction

63

3.1.3

On Overcoming the Issues of Knowledge Extraction Approaches

68

3.1.4

The Future of Math Embeddings

70

3.2

Semantification with Mathematical Objects of Interest

70

3.2.1

Related Work

72

3.2.2

Data Preparation

72

3.2.3

Frequency Distributions of Mathematical Formulae

76

3.2.4

Relevance Ranking for Formulae

81

3.2.5

Applications

87

3.2.6

Outlook

91

3.3

Semantification with Textual Context Analysis

91

3.3.1

Semantification, Translation & Evaluation Pipeline

91

CHAPTER 4

From LaTeX to Computer Algebra Systems

95

4.1

Context-Agnostic Neural Machine Translation

96

4.1.1

Training Datasets & Preprocessing

96

4.1.2

Methodology

97

4.1.3

Evaluation of the Convolutional Network

97

4.2

Context-Sensitive Translation

101

4.2.1

Motivation

101

4.2.2

Related Work

104

4.2.3

Formal Mathematical Language Translations

104

4.2.4

Document Pre-Processing

108

4.2.5

Annotated Dependency Graph Construction

108

4.2.6

Semantic Macro Replacement Patterns

110

CHAPTER 5

Qualitative and Quantitative Evaluations

113

5.1

Evaluations on the Digital Library of Mathematical Functions

114

5.1.1

The DLMF dataset

116

5.1.2

Semantic LaTeX to CAS translation

117

5.1.3

Evaluation of the DLMF using CAS

123

5.1.4

Results

128

5.1.5

Conclude Quantitative Evaluations on the DLMF

131

5.2

Evaluations on Wikipedia

132

5.2.1

Symbolic and Numeric Testing

133

5.2.2

Benchmark Testing

133

5.2.3

Results

134

5.2.4

Error Analysis & Discussion

137

5.2.5

Conclude Qualitative Evaluations on Wikipedia

139

CHAPTER 6

Conclusion and Future Work	141
6.1 Summary	141
6.2 Contributions and Impact of the Thesis	150
6.3 Future Work	153
6.3.1 Improved Translation Pipeline	154
6.3.2 Improve LaTeX to MathML Converters	155
6.3.3 Enhanced Formulae in Wikipedia	156
6.3.4 Language Independence	158
BACK MATTER	161
Glossary	161
Bibliography of Publications, Submissions & Talks	171
Bibliography	173