## Contents

### List of Figures 9

### List of Tables 11

### 1 Introduction 15

#### 1.1 Overview on the Biomedical Relevance and Information Resources of Chemical Entities and Histone Modifications 17

- 1.1.1 Introduction to Epigenetics and its Role in Biology, Medicine, and Pharmacology 19

#### 1.2 Overview on Text Processing Methods 22

- 1.2.1 Introduction to Information Extraction 23
  - 1.2.1.1 Introduction to Named Entity Recognition 25

- 1.2.2 Challenges of Chemical Named Entity and Histone Modification Term Recognition 26
  - 1.2.2.1 Overview on Terminology of Chemical Substances 27
  - 1.2.2.2 Overview on Designators of Chemical Named Entities 30
  - 1.2.2.3 Overview on Histone Modification Terminology 32

### 2 Problem Description and Goal 35

#### 2.1 Outline of the Thesis 36

### I Fundamentals 39

### 3 Fundamentals on Applied and Developed Methods 41

#### 3.1 Information Extraction Techniques 41

- 3.1.1 Overview on Methods applied in IE 41

- 3.1.2 Named Entity Recognition 42
  - 3.1.2.1 Literature Survey on Biomedical and Chemical Named Entity Recognition 42
  - 3.1.2.2 Dictionary-based NER Approaches 44
  - 3.1.2.3 Machine Learning-based NER Approaches 49

- 3.1.3 Term Normalization 52
  - 3.1.3.1 Generation of Canonical Term Representatives 53
  - 3.1.3.2 Mapping of Terms to Reference Identifiers 53

- 3.1.4 Corpus Selection and Annotation 55
  - 3.1.4.1 Active Learning 56

- 3.1.5 Evaluation Measures 57
3.2 Function Annotation of Entities ........................................ 59
  3.2.1 Impact of Ontologies for Function Annotation and Data Management 59
  3.2.2 Information Extraction for Supporting Function Annotation of Entities 60
    3.2.2.1 Impact of Ontology for Information Extraction .............. 61
    3.2.2.2 Ontology Learning ........................................ 61
  3.2.3 Function Prediction of Chemical Entities .......................... 66
    3.2.3.1 Related Work – Class Prediction of Chemical Compounds . 67
    3.2.3.2 Classification Methods .................................. 68
  3.3 Data Visualization .................................................... 70

II Developed Systems and Implementations ................................ 75

4 Building a Framework for the Information Aggregation of Chemical Entities 77
  4.1 Recognition of Chemical Named Entities in Text .................... 77
    4.1.1 Generation of an Evaluation Text Corpus and Annotation of Chemical Entity Classes ........................................ 77
    4.1.2 Generation of a Chemical Named Entity Dictionary .......... 80
      4.1.2.1 Raw Dictionary Generation and Performance Analysis .... 81
      4.1.2.2 Improvement of the Dictionary Quality by Curation ..... 88
      4.1.2.3 Adjusting the Approximate String Matching to the Chemical Domain .............................................. 92
      4.1.2.4 Evaluation of the Dictionary Curation and the Approximate String Matching ....................................... 93
      4.1.2.5 Generation of a Final Chemical Dictionary ............... 95
    4.1.3 Results of the Chemical Named Entity Recognition ............ 121
    4.1.4 Application of Recognized Chemical Named Entities for Information Retrieval ............................................. 125
    4.1.5 Summary of the Chemical Named Entity Recognition and Discussion 128
  4.2 Extraction of Function Annotation Information on Chemical Compounds from Text .............................................. 132
    4.2.1 Developed Methodology for the Extraction of Function Annotation Information on Chemical Entities from Text .................. 133
      4.2.1.1 Description of A New Method for the Extraction of Property Information on Chemical Compounds from Text ........... 133
      4.2.1.2 Generation of Evaluation Corpora .......................... 137
      4.2.1.3 Evaluation of the Hypernymic Phrase Extraction .......... 138
      4.2.1.4 Results of the Function Annotation Term Extraction Procedure 139
      4.2.1.5 Summary of the Developed Workflow for the Extraction of Function Annotation Terms and Discussion of the Results .... 140
  4.3 Developed Framework for Information Aggregation and Annotation of Chemical Compounds ............................................ 142
    4.3.1 Description of the Workflow for Finding New Function Annotations of Chemical Entities in Text ............................ 142
      4.3.1.1 Results of the Term Comparison Procedure ................ 145
4.3.2 Application of the Information Aggregation Framework: Drug Classification Schema Extension ........................................... 148
  4.3.2.1 Description of a New Method for the Extension of Classification Schemes by New Drug Instances ......................... 149
  4.3.2.2 Results of the Classification Schema Extension ................. 152
4.3.3 Summary of the Developed Information Aggregation Framework and Discussion of the Results ................................. 153

5 Histone Modification Recognition in Text ........................................ 157
  5.1 Development of a CRF-based NER Approach for Recognizing Histone Modifications ...................................................... 157
    5.1.1 Corpus Selection and Annotation ........................................ 158
      5.1.1.1 Corpus Generation ............................................... 158
      5.1.1.2 Corpus Annotation ............................................. 158
    5.1.2 CRF Training and Feature Selection .................................... 159
      5.1.2.1 Feature Selection ............................................. 160
      5.1.2.2 Recognition Problems ......................................... 162
    5.1.3 Improvement of a primary CRF Model by the Extension of the Training Corpus through Active Learning .................. 163
  5.2 Canonicalization of Histone Modification Terms .......................... 165
    5.2.1 Canonicalization Workflow .......................................... 165
    5.2.2 Evaluation of the Canonicalization ..................................... 168
  5.3 Development of a Dictionary-based Approach for Recognizing Histone Modification Terms in Text ................................. 170
    5.3.1 Generation of a Histone Modification Term Dictionary ............ 170
    5.3.2 Comparison of the Results from ProMiner_{Hmod} and the CRF Approach .......................................................... 171
  5.4 Generation of a Histone Modification Concept Hierarchy ............... 173
    5.4.1 Automated Support of the Hierarchy Extension ...................... 175
  5.5 Applications of the Histone Modification Identification Approach and the Concept Hierarchy .................................... 176
    5.5.1 Analysis of the Information Content of Histone Modification Descriptions Extracted from MEDLINE ......................... 176
    5.5.2 Document Retrieval ................................................. 177
    5.5.3 Application of the Histone Modification Finding Results .......... 179
  5.6 Summary and Discussion ...................................................... 188

6 Conclusion and Perspectives .......................................................... 191
  6.1 Conclusion ............................................................................. 191
  6.2 Perspectives for Future Research .............................................. 193

A Definitions and Data Resources ...................................................... 197
  A.1 Definitions ............................................................................ 197
  A.2 Data Resources ..................................................................... 198
    A.2.1 Data Sources on Chemical Compounds ................................ 198
    A.2.2 Pharmacological Classification Schemes on Chemical Compounds . 199