Relation extraction methods for biomedical literature

Bui, Q.C.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Contents

1. Introduction 1
  1.1 Motivation 1
  1.2 Research questions 3
  1.3 Outline of the thesis 4

2. Relation extraction methods for biomedical text 5
  2.1 Introduction 5
    2.1.1 Text preprocessing 6
    2.1.2 Named entity recognition 7
    2.1.3 Parsing 8
    2.1.4 Relation extraction 8
    2.1.5 Evaluation metrics 8
  2.2 Natural Language Processing 9
    2.2.1 Sentence splitting 9
    2.2.2 Lexical processing 9
    2.2.3 Syntactic processing 10
  2.3 Machine learning 12
    2.3.1 SVM methods for relation extraction 14
    2.3.2 Kernel methods 15
  2.4 Biomedical corpora 15
    2.4.1 GENIA corpus 15
    2.4.2 PPI corpora 15
    2.4.3 GENIA events corpus 16
  2.5 Relation extraction methods 17
    2.5.1 Co-occurrence approaches 17
    2.5.2 Pattern-based approaches 18
    2.5.3 Rule-based approaches 19
    2.5.4 Machine learning-based approaches 20
    2.5.5 Performance comparison of existing approaches to relation extraction 22
  2.6 Conclusion 23

3. Extracting causal relations on HIV drug resistance from literature 25
  3.1 Introduction 26
  3.2 System and methods 27
    3.2.1 Text retrieval 28
    3.2.2 Text preprocessing 29
    3.2.3 Relation extraction 32
    3.2.4 Relation combination 34
  3.3 Results and Discussion 37
    3.3.1 Datasets 37
    3.3.2 Relation extraction performance 37
    3.3.3 Relation combination performance 39
  3.4 Conclusions 42

4. A hybrid approach to extract protein-protein interactions 43
  4.1 Introduction 44
# Contents

4.2 System and methods 45  
4.2.1 Text preprocessing 46  
4.2.2 Extracting candidate PPI pairs 46  
4.3 Results and Discussion 52  
4.3.1 Data sets 52  
4.3.2 Evaluation methods 52  
4.3.3 Performance of PPI extraction algorithm 53  
4.3.4 Single corpus evaluation 54  
4.3.5 Cross-corpora evaluation 56  
4.3.6 Performance time 58  
4.4 Conclusions 58  

5. A robust approach to extract biomedical events from literature 61  
5.1 Introduction 62  
5.2 System and methods 64  
5.2.1 Structured representation 64  
5.2.2 Learning rules 65  
5.2.3 Rule combination 68  
5.2.4 Event extraction 70  
5.2.5 Text preprocessing 72  
5.3 Results and discussion 72  
5.3.1 Datasets 72  
5.3.2 Evaluation settings 73  
5.3.3 Event extraction 73  
5.3.4 Performance time 76  
5.3.5 Performance analysis 77  
5.4 Conclusion 79  

6. Discussion and Conclusion 81  
6.1 The use of syntactic information for different relation extraction tasks 81  
6.2 The role of machine learning to relation extraction tasks 82  
6.3 The performance time of relation extraction systems 83  
6.4 Contribution 84  
6.5 Future work 85  

References 87  
Summary 97  
Samenvatting 99  
List of publications 101  
Acknowledgements 103