## Contents

List of Figures .......................................................... xvii
List of Tables ............................................................ xix

1 Introduction ..........................................................  1
   1.1 The Problem: Humor and Irony .................................  3
   1.2 The Core of the Problem: (Figurative) Language in Social Media 4
   1.3 Objective ..........................................................  6
   1.4 Thesis Overview ..................................................  8

2 Figurative Language .................................................. 11
   2.1 Background ........................................................ 11
   2.2 Literal Language .................................................. 13
   2.3 Figurative Language ................................................ 16
   2.4 Figurative Devices ................................................. 21
      2.4.1 Metaphor ...................................................... 22
      2.4.2 Metonym ...................................................... 23
      2.4.3 Simile ......................................................... 24
      2.4.4 Idioms ......................................................... 25
   2.5 Figurative Devices in this Thesis ............................... 26
      2.5.1 Humor: A Multidimensional Phenomenon ................. 26
      2.5.2 Irony: A Veiled Phenomenon ............................... 30
   2.6 Summary ........................................................... 33
<table>
<thead>
<tr>
<th>Contents</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 Figurative Language Processing</td>
<td>35</td>
</tr>
<tr>
<td>3.1 Natural Language Processing</td>
<td>35</td>
</tr>
<tr>
<td>3.2 Figurative Language Processing</td>
<td>37</td>
</tr>
<tr>
<td>3.3 Advances on FLP</td>
<td>39</td>
</tr>
<tr>
<td>3.3.1 Metaphor Processing</td>
<td>40</td>
</tr>
<tr>
<td>3.3.2 Metonym Processing</td>
<td>40</td>
</tr>
<tr>
<td>3.3.3 Similes Processing</td>
<td>41</td>
</tr>
<tr>
<td>3.3.4 Idioms Processing</td>
<td>42</td>
</tr>
<tr>
<td>3.4 Related Work on Humor Processing</td>
<td>43</td>
</tr>
<tr>
<td>3.4.1 Humor Generation</td>
<td>43</td>
</tr>
<tr>
<td>3.4.2 Humor Recognition</td>
<td>45</td>
</tr>
<tr>
<td>3.5 Related Work on Irony Processing</td>
<td>47</td>
</tr>
<tr>
<td>3.5.1 Irony Detection</td>
<td>48</td>
</tr>
<tr>
<td>3.5.2 Sarcasm and Satire Detection</td>
<td>49</td>
</tr>
<tr>
<td>3.6 Summary</td>
<td>50</td>
</tr>
<tr>
<td>4 Automatic Humor Recognition</td>
<td>53</td>
</tr>
<tr>
<td>4.1 Initial Assumptions</td>
<td>53</td>
</tr>
<tr>
<td>4.2 Humor Recognition Model</td>
<td>56</td>
</tr>
<tr>
<td>4.2.1 Ambiguity</td>
<td>56</td>
</tr>
<tr>
<td>4.2.2 Lexical Ambiguity</td>
<td>57</td>
</tr>
<tr>
<td>4.2.3 Morphological Ambiguity</td>
<td>58</td>
</tr>
<tr>
<td>4.2.4 Syntactic Ambiguity</td>
<td>59</td>
</tr>
<tr>
<td>4.2.5 Semantic Ambiguity</td>
<td>59</td>
</tr>
<tr>
<td>4.3 Evaluation of Ambiguity-based Patterns</td>
<td>60</td>
</tr>
<tr>
<td>4.3.1 Data Sets H1 - H3</td>
<td>60</td>
</tr>
<tr>
<td>4.3.2 Evaluation</td>
<td>61</td>
</tr>
<tr>
<td>4.3.2.1 Lexical Layer: Perplexity</td>
<td>62</td>
</tr>
<tr>
<td>4.3.2.2 Morphological Layer: POS Tags</td>
<td>63</td>
</tr>
<tr>
<td>4.3.2.3 Syntactic Layer: Sentence Complexity</td>
<td>66</td>
</tr>
<tr>
<td>4.3.2.4 Semantic Layer: Sense Dispersion</td>
<td>67</td>
</tr>
<tr>
<td>4.3.3 Discussion of Preliminary Findings</td>
<td>69</td>
</tr>
<tr>
<td>4.4 Adding Surface Patterns</td>
<td>70</td>
</tr>
</tbody>
</table>
## 4.4.1 Humor Domain

Page 71

## 4.4.2 Polarity

Page 72

## 4.4.3 Templates

Page 72

## 4.4.4 Affectiveness

Page 73

## 4.5 HRM Evaluation

Page 73

### 4.5.1 Data Set H4

Page 74

### 4.5.2 Evaluation

Page 75

### 4.5.3 Results and Overall Discussion

Page 77

## 4.6 Summary

Page 80

## 5 Irony Detection

### 5.1 Irony: Beyond a Funny Effect

Page 84

### 5.2 Target

Page 85

### 5.3 Basic Irony Detection Model

Page 86

#### 5.3.1 Data Set I1

Page 87

#### 5.3.2 HRM and Irony

Page 89

### 5.4 Basic IDM representation

Page 90

#### 5.4.1 N-grams

Page 90

#### 5.4.2 Descriptors

Page 91

#### 5.4.3 POS n-grams

Page 92

#### 5.4.4 Funniness

Page 93

#### 5.4.5 Polarity

Page 94

#### 5.4.6 Affectiveness

Page 94

#### 5.4.7 Pleasantness

Page 94

### 5.5 Evaluation

Page 95

#### 5.5.1 Discussion

Page 96

#### 5.5.2 Pattern Analysis

Page 99

### 5.6 Complex Irony Detection Model

Page 100

#### 5.6.1 Signatures

Page 101

#### 5.6.2 Unexpectedness

Page 102

#### 5.6.3 Style

Page 103

#### 5.6.4 Emotional Contexts

Page 104

#### 5.6.5 Data Set I2

Page 105
CONTENTS

B  Examples of Blogs in H4  159

C  Set of Features in Signatures  163

D  Examples of Patterns Regarding the Complex Irony Detection Model  165

E  Probability Density Function for Patterns in Complex Irony Detection Model  169

F  Examples of the Most Ironic Sentences  175
List of Figures

2.1 Example of visual humor†. ........................................... 28

4.1 Frequency of Internet searches related to 5 different social subjects during four years (September 2007 - 2011) around the world. Statistics retrieved from Google Insights. ...................... 54

4.2 Probability of assigning POS tags concerning positive and negative examples in H1 - H3. ........................................ 65

4.3 Chaos parser: Example of syntactic representation. .............. 66

4.4 Humor average per subset. .......................................... 76

5.1 Representativeness ratios of patterns funny (a), polarity (b) (positive in red and negative in blue, respectively), affectiveness (c), and pleasantness (d). Axis x represents the ironic reviews whereas axis y depicts its representativeness ratio. Dotted line symbolizes representativeness threshold. ........................................ 96

5.2 Learning curve according to sets AMA (a), SLA (b), and TRI (c). 98

5.3 Distribution of positive, negative and out of vocabulary (neutral) terms in I2. ...................................................... 111

5.4 Classification accuracy regarding irony vs. education (a), humor (b), and politics (c), considering a balanced distribution. .... 113

5.5 Classification accuracy regarding irony vs. education (a), humor (b), and politics (c), considering an imbalanced distribution. ... 113

5.6 Relevance of every single dimension according to its information gain value. ...................................................... 115

6.1 γ values per set: movies2 (a); movies1 (b); books (c); articles (d). 123
LIST OF FIGURES

6.2 Preliminary humor taxonomy ............................................. 134

E.1 Probability density function for dimensions in signatures. ........ 170
E.2 Probability density function for dimensions in unexpectedness. . 171
E.3 Probability density function for dimensions in style. ............... 172
E.4 Probability density function for dimensions in emotional contexts. 173
List of Tables

4.1 Detailed information regarding data sets H1 to H3. ............ 62
4.2 Perplexity ratios. ........................................... 64
4.3 Average of POS tags per example. .............................. 64
4.4 Results concerning sentence complexity. ......................... 67
4.5 Semantic dispersion at sentence and word level. ............... 68
4.6 H4 characteristics. Measures: corpus vocabulary size (CVS); doc-
ument and vocabulary length (DL and VL, respectively); vocabu-
lary and document length ratio (VDR); unsupervised vocabulary
based measure (UVB); stylometric evaluation measure (SEM). 75
4.7 Results obtained with NB classifier. ............................ 78
4.8 Results obtained with SVM classifier. ............................ 78
4.9 Information gain results on HRM’s patterns. .................... 79

5.1 Detailed information regarding data set I1. ....................... 89
5.2 Results applying HRM over I1. ................................. 90
5.3 Statistics of the most frequent word n-grams. .................. 91
5.4 SenseCluster parameters per experiment. ........................ 92
5.5 Descriptors obtained with keyness and clustering metrics. .... 92
5.6 Statistics of the most frequent POS-grams. ..................... 93
5.7 Classification results. ........................................... 97
5.8 Most discriminating patterns per set. ........................... 98
5.9 Statistics in terms of tokens per set concerning data set I2. ... 107
5.10 Monge Elkan distance among sets. ............................. 108
5.11 Overall pattern representativeness per set. ..................... 110
5.12 Semantic relatedness per set. .................................. 111
LIST OF TABLES

5.13 Precision, Recall and F-Measure regarding i) balanced distribution, and ii) imbalanced distribution. ........................................... 114

6.1 Classification accuracy of funny vs. informative ($c_1$), insightful ($c_2$), and negative ($c_3$), respectively. ........................................... 119

6.2 Manual evaluation in terms of isolated sentences. ......................................................... 125

6.3 Manual evaluation in terms of whole documents. ......................................................... 125

6.4 Statistics regarding annotators judgments. ................................................................. 130

6.5 Irony retrieval results. ................................................................................................. 132

C.1 Features in pattern signatures. .................................................................................. 164