

Contents

Part I Background

1	Introduction	3
1.1	Semantic Hierarchy of Minimal Propositions through Discourse-Aware Sentence Splitting	4
1.2	Formal Semantic Representation through Open Information Extraction	9
1.3	Problems in Syntactic Text Simplification and Open Information Extraction	10
1.4	Research Questions	13
1.5	Contributions	15
1.6	Outline	16
1.7	Associated Publications	17
2	Related Work	21
2.1	Text Simplification	21
2.1.1	Approaches to Simplification	22
2.1.1.1	Rule-based Approaches	22
2.1.1.2	Approaches based on Monolingual Statistical Machine Translation	27
2.1.1.3	Approaches based on the Induction of Synchronous Grammars	29
2.1.1.4	Semantics-Assisted Approaches	30
2.1.1.5	Neural Sequence-to-Sequence Approaches	35
2.1.1.6	Summary	40
2.1.2	Data Resources for Simplification	44

	2.1.2.1	Text Simplification Corpora	44
	2.1.2.2	Sentence Splitting Corpora	51
	2.1.2.3	Summary	54
2.2		Text Coherence	55
	2.2.1	Discourse-level Syntactic Simplification	55
	2.2.2	Discourse Parsing	56
2.3		Open Information Extraction	59
	2.3.1	Approaches to Open Information Extraction	60
	2.3.1.1	Early Approaches based on Hand-crafted Extraction Patterns and Self-supervised Learning	60
	2.3.1.2	Paraphrase-based Approaches	65
	2.3.1.3	Systems that Capture Inter-Proposition Relationships	68
	2.3.1.4	End-to-End Neural Approaches	71
	2.3.2	Benchmarking Open Information Extraction Approaches	72
2.4		Meaning Representations	81

Part II Discourse-Aware Sentence Splitting

3	Introduction	87
4	Subtask 1: Splitting into Minimal Propositions	91
	4.1 Property of Minimality	92
	4.1.1 Minimality on the Syntactic Level	92
	4.1.2 Minimality on the Semantic Level	96
	4.2 Splitting Procedure	98
	4.3 Execution Order of the Transformation Patterns	101
5	Subtask 2: Establishing a Semantic Hierarchy	105
	5.1 Constituency Type Classification	105
	5.2 Rhetorical Relation Identification	107
6	Transformation Patterns	111
	6.1 Clausal Disembedding	112
	6.1.1 Coordinate Clauses	112
	6.1.2 Adverbial Clauses	113
	6.1.3 Relative Clauses	115
	6.1.4 Reported Speech	124
	6.2 Phrasal Disembedding	127

6.2.1	Coordinate Verb Phrases	128
6.2.2	Coordinate Noun Phrase Lists	129
6.2.3	Participial Phrases	130
6.2.4	Appositive Phrases	132
6.2.5	Prepositional Phrases	138
6.2.6	Adjectival and Adverbial Phrases	139
6.2.7	Lead Noun Phrases	140
7	Transformation Process	143
7.1	Data Model: Linked Proposition Tree	143
7.2	Transformation Algorithm	145
7.3	Transformation Example	148
8	Sentence Splitting Corpus	161
9	Summary	165
 Part III Open Information Extraction		
10	Introduction	169
11	Subtask 3: Extracting Semantically Typed Relational Tuples	171
11.1	Enriching State-of-the-Art Open Information Extraction Approaches with Semantic Information	172
11.2	Reference Implementation Graphene	176
12	Generating a Formal Semantic Representation	179
12.1	Lightweight Semantic Representation for Open Information Extraction	179
12.2	Human Readable Representation	183
12.3	Machine Readable Representation	185
13	Summary	189
 Part IV Evaluation		
14	Experimental Setup	193
14.1	Subtask 1: Splitting into Minimal Propositions	193
14.1.1	Datasets	194
14.1.2	Baselines	194
14.1.3	Automatic Metrics	195
14.1.3.1	Descriptive Statistics	195
14.1.3.2	Translation Assessment	195

14.1.3.3	Syntactic Complexity	196
14.1.3.4	Rule Application Statistics	196
14.1.4	Manual Analyses	196
14.1.4.1	Human Judgments	197
14.1.4.2	Qualitative Analysis of the Transformation Patterns	197
14.2	Subtask 2: Establishing a Semantic Hierarchy	199
14.2.1	Automatic Metrics	199
14.2.1.1	Constituency Type Classification	200
14.2.1.2	Rhetorical Relation Identification	200
14.2.2	Manual Analysis	201
14.3	Subtask 3: Extracting Relations and their Arguments	202
14.3.1	Baselines	203
14.3.2	Benchmarks	204
14.3.3	Comparative Analysis of the Outputs	204
14.3.3.1	Performance of the Reference Open Information Extraction Implementation Graphene	205
14.3.3.2	Sentence Splitting as a Pre-processing Step	209
14.3.4	Manual Analyses	210
14.3.4.1	Error Analysis	211
14.3.4.2	Qualitative Analysis of the Extracted Relations	211
14.3.5	Analysis of the Lightweight Semantic Representation of Relational Tuples	211
14.4	Sentence Splitting Corpus	212
14.4.1	Automatic Metrics	212
14.4.2	Manual Analysis	212
15	Results and Discussion	213
15.1	Subtask 1: Splitting into Minimal Propositions	214
15.1.1	Automatic Metrics	214
15.1.1.1	WikiLarge Corpus	214
15.1.1.2	Newsela Corpus	217
15.1.1.3	WikiSplit Corpus	217
15.1.1.4	Rule Application Statistics	218
15.1.2	Manual Analyses	220
15.1.2.1	Human Judgments	220

15.1.2.2	Qualitative Analysis of the Transformation Patterns	222
15.2	Subtask 2: Establishing a Semantic Hierarchy	226
15.2.1	Automatic Metrics	226
15.2.1.1	Constituency Type Classification	227
15.2.1.2	Rhetorical Relation Identification	230
15.2.2	Manual Analysis	235
15.3	Subtask 3: Extracting Relations and their Arguments	238
15.3.1	Performance of the Reference Open Information Extraction Implementation Graphene	238
15.3.1.1	Automatic Comparative Analysis of the Outputs	239
15.3.1.2	Error Analysis	248
15.3.1.3	Qualitative Analysis of the Extracted Relations	250
15.3.2	Sentence Splitting as a Pre-processing Step	254
15.3.2.1	Automatic Comparative Analysis of the Outputs	255
15.3.2.2	Qualitative Analysis of the Extracted Relations	262
15.3.2.3	Analysis of the Lightweight Semantic Representation of Relational Tuples	268
15.4	Sentence Splitting Corpus	272
15.4.1	Automatic Metrics	272
15.4.2	Manual Analysis	273
 Part V Conclusion		
16	Summary and Contributions	277
16.1	Chapter I	277
16.2	Chapter II	278
16.3	Chapter III	279
16.4	Chapter IV	280
16.5	Chapter V	281
17	Discussion and Conclusions	283
17.1	Semantic Hierarchy of Minimal Propositions	283
17.1.1	Subtask 1: Splitting into Minimal Propositions	283
17.1.1.1	Hypothesis 1.1: Syntactic Correctness	284
17.1.1.2	Hypothesis 1.2: Minimality	285

17.1.1.3	Hypothesis 1.3: Semantic Correctness	287
17.1.1.4	Summary	288
17.1.2	Subtask 2: Establishing a Semantic Hierarchy	288
17.1.2.1	Hypothesis 1.4: Contextual Hierarchy	289
17.1.2.2	Hypothesis 1.5: Coherence Structure	290
17.1.2.3	Summary	291
17.2	Semantically Typed Relational Tuples	291
17.2.1	Subtask 3: Extracting Relations and their Arguments	292
17.2.1.1	Hypothesis 2.1: Coverage and Accuracy	292
17.2.1.2	Hypothesis 2.2: Canonical Context-Preserving Representation of Relational Tuples	294
17.2.2	Summary	295
18	Limitations and Future Research Directions	297
	Bibliography	301