

# CONTENTS

---

1	INTRODUCTION	1
1.1	Background: Identity in the digital era . . . . .	1
1.2	Challenge: Entity Linking in the long tail . . . . .	4
1.3	Research questions . . . . .	8
1.4	Approach and structure of the thesis . . . . .	9
1.4.1	Describing and observing the head and the tail . . . . .	10
1.4.2	Analyzing the evaluation bias on the long tail . . . . .	11
1.4.3	Improving the evaluation bias on the long tail . . . . .	11
1.4.4	Enabling access to knowledge about long-tail entities beyond DBpedia . . . . .	12
1.4.5	The role of knowledge in establishing identity of long-tail entities . . . . .	13
1.5	Summary of findings . . . . .	13
1.6	Software and data . . . . .	14
2	DESCRIBING AND OBSERVING THE HEAD AND THE TAIL OF ENTITY LINKING	17
2.1	Introduction . . . . .	17
2.2	Related work . . . . .	18
2.3	Approach . . . . .	20
2.3.1	The head-tail phenomena of the entity linking task . . . . .	20
2.3.2	Hypotheses on the head-tail phenomena of the entity linking task . . . . .	23
2.3.3	Datasets and systems . . . . .	25
2.3.4	Evaluation . . . . .	25
2.4	Analysis of data properties . . . . .	26
2.4.1	Frequency distribution of forms and instances in datasets . . . . .	26
2.4.2	PageRank distribution of instances in datasets . . . . .	27
2.4.3	Ambiguity distribution of forms . . . . .	27
2.4.4	Variance distribution of instances . . . . .	28
2.4.5	Interaction between frequency, PageRank, and ambiguity/-variance . . . . .	28
2.4.6	Frequency distribution for a single form or an instance . . . . .	30
2.5	Analysis of system performance and data properties . . . . .	32
2.5.1	Correlating system performance with form ambiguity . . . . .	32
2.5.2	Correlating system performance with form frequency, instance frequency, and PageRank . . . . .	33
2.5.3	Correlating system performance with ambiguity and frequency of forms jointly . . . . .	34
2.5.4	Correlating system performance with frequency of instances for ambiguous forms . . . . .	35
2.6	Summary of findings . . . . .	37

2.7	Recommended actions . . . . .	37
2.8	Conclusions . . . . .	38
3	ANALYZING THE EVALUATION BIAS ON THE LONG TAIL OF DISAM- BIGUATION & REFERENCE . . . . .	41
3.1	Introduction . . . . .	42
3.2	Temporal aspect of the disambiguation task . . . . .	43
3.3	Related work . . . . .	46
3.4	Preliminary study of EL evaluation datasets . . . . .	47
3.4.1	Datasets . . . . .	47
3.4.2	Dataset characteristics . . . . .	49
3.4.3	Distributions of instances and surface forms . . . . .	52
3.4.4	Discussion and roadmap . . . . .	56
3.5	Semiotic generation and context model . . . . .	58
3.6	Methodology . . . . .	59
3.6.1	Metrics . . . . .	59
3.6.2	Tasks . . . . .	62
3.6.3	Datasets . . . . .	63
3.7	Analysis . . . . .	64
3.8	Proposal for improving evaluation . . . . .	67
3.9	Conclusions . . . . .	68
4	IMPROVING THE EVALUATION BIAS ON THE LONG TAIL OF DISAM- BIGUATION & REFERENCE . . . . .	71
4.1	Introduction . . . . .	72
4.2	Motivation & target communities . . . . .	74
4.2.1	Disambiguation & reference . . . . .	74
4.2.2	Reading Comprehension & Question Answering . . . . .	75
4.2.3	Moving away from semantic overfitting . . . . .	76
4.3	Task requirements . . . . .	76
4.4	Methods for creating an event-based task . . . . .	77
4.4.1	State of text-to-data datasets . . . . .	77
4.4.2	From data to text . . . . .	80
4.5	Data & resources . . . . .	84
4.5.1	Structured data . . . . .	84
4.5.2	Example document . . . . .	85
4.5.3	Licensing & availability . . . . .	85
4.6	Task design . . . . .	86
4.6.1	Subtasks . . . . .	86
4.6.2	Question template . . . . .	86
4.6.3	Question creation . . . . .	87
4.6.4	Data partitioning . . . . .	89
4.7	Mention annotation . . . . .	89
4.7.1	Annotation task and guidelines . . . . .	90
4.7.2	Annotation environment . . . . .	91
4.7.3	Annotation process . . . . .	93
4.7.4	Corpus description . . . . .	93

4.8	Evaluation . . . . .	96
4.8.1	Criteria . . . . .	96
4.8.2	Baselines . . . . .	96
4.9	Participants . . . . .	97
4.10	Results . . . . .	98
4.10.1	Incident-level evaluation . . . . .	98
4.10.2	Document-level evaluation . . . . .	100
4.10.3	Mention-level evaluation . . . . .	102
4.11	Discussion . . . . .	102
4.12	Conclusions . . . . .	103
5	ENABLING ACCESS TO KNOWLEDGE ON THE LONG-TAIL ENTITIES BEYOND DBPEDIA . . . . .	107
5.1	Introduction . . . . .	108
5.2	Problem description . . . . .	110
5.2.1	Requirements . . . . .	110
5.2.2	Current state-of-the-art . . . . .	111
5.3	Related work . . . . .	113
5.4	Access to entities at LOD scale with LOD Lab . . . . .	115
5.4.1	LOD Lab . . . . .	115
5.4.2	APIs and tools . . . . .	117
5.5	LOTUS . . . . .	118
5.5.1	Model . . . . .	118
5.5.2	Language tags . . . . .	119
5.5.3	Linguistic entry point to the LOD Cloud . . . . .	121
5.5.4	Retrieval . . . . .	121
5.6	Implementation . . . . .	123
5.6.1	System architecture . . . . .	123
5.6.2	Implementation of the matching and ranking algorithms . . . . .	125
5.6.3	Distributed architecture . . . . .	126
5.6.4	API . . . . .	126
5.6.5	Examples . . . . .	127
5.7	Performance statistics and flexibility of retrieval . . . . .	128
5.7.1	Performance statistics . . . . .	128
5.7.2	Flexibility of retrieval . . . . .	130
5.8	Finding entities beyond DBpedia . . . . .	130
5.8.1	AIDA-YAGO2 . . . . .	131
5.8.2	Local monuments guided walks . . . . .	132
5.8.3	Scientific journals . . . . .	133
5.9	Discussion and conclusions . . . . .	134
6	THE ROLE OF KNOWLEDGE IN ESTABLISHING IDENTITY OF LONG- TAIL ENTITIES . . . . .	137
6.1	Introduction . . . . .	138
6.2	Related work . . . . .	140
6.2.1	Entity Linking and NIL clustering . . . . .	140
6.2.2	Attribute extraction . . . . .	142

6.2.3	Knowledge Base Completion (KBC)	142
6.2.4	Other knowledge completion variants	143
6.3	Task and hypotheses	144
6.3.1	The NIL clustering task	144
6.3.2	Research question and hypotheses	145
6.4	Profiling	146
6.4.1	Aspects of profiles	146
6.4.2	Examples	147
6.4.3	Definition of a profile	148
6.4.4	Neural methods for profiling	149
6.5	Experimental setup	151
6.5.1	End-to-end pipeline	151
6.5.2	Data	152
6.5.3	Evaluation	154
6.5.4	Automatic attribute extraction	155
6.5.5	Reasoners	158
6.6	Extrinsic evaluation	159
6.6.1	Using explicit information to establish identity	159
6.6.2	Profiling implicit information	161
6.6.3	Analysis of ambiguity	166
6.7	Intrinsic analysis of the profiler	166
6.7.1	Comparison against factual data	166
6.7.2	Comparison against human expectations	169
6.8	Discussion and limitations	170
6.8.1	Summary of the results	170
6.8.2	Harmonizing knowledge between text and knowledge bases	171
6.8.3	Limitations of profiling by NNs	172
6.9	Conclusions and future work	172
7	CONCLUSION	179
7.1	Summarizing our results	179
7.1.1	Describing and observing the head and the tail of Entity Linking	179
7.1.2	Analyzing the evaluation bias on the long tail	180
7.1.3	Improving the evaluation on the long tail	181
7.1.4	Enabling access to knowledge on the long-tail entities	182
7.1.5	The role of knowledge in establishing identity of long-tail entities	183
7.2	Lessons learned	184
7.2.1	Observations	184
7.2.2	Recommendations	185
7.3	Future research directions	187
7.3.1	Engineering of systems	187
7.3.2	Novel tasks	188
7.3.3	A broader vision for the long tail	189

BIBLIOGRAPHY

191