

Contents

1	Introduction to NLG	1
1.1	What Is Natural Language Generation?	2
1.2	Example: Weather Forecast (<i>Data-to-Text</i>)	3
1.2.1	Use Case: Point Weather Forecasts for General Public	4
1.2.2	Technology: Rule-Based NLG	5
1.2.3	Evaluation	6
1.3	Example: Summarising Consultations (<i>Text-to-Text</i>)	6
1.3.1	Use Case: Generating Summaries of Doctor-Patient Consultations	6
1.3.2	Technology	8
1.3.3	Evaluation	8
1.4	Technologies	9
1.4.1	Rule-Based NLG	10
1.4.2	Machine Learning and Neural Models	11
1.4.3	Combining Rules and ML	12
1.5	Effectiveness	13
1.5.1	Requirements	13
1.5.2	Evaluation	14
1.5.3	Safety, Testing, and Maintenance	15
1.6	Use Cases and Applications	16
1.7	Ethics	18
1.8	A Very Short History of NLG	19
1.8.1	Early History	19
1.8.2	1990–2014	20
1.8.3	2015–2024	20
1.8.4	My Personal NLG Journey	22
1.9	Resources and Further Reading	22
2	Rule-Based NLG	25
2.1	NLG Pipeline	26
2.2	Examples	27

2.2.1	Driving Feedback	28
2.2.2	Babytalk	29
2.3	Signal Analysis	30
2.3.1	Noise Detection: Principles	31
2.3.2	Pattern Detection: Principles	31
2.3.3	Techniques for Signal Analysis	33
2.4	Data Interpretation	33
2.4.1	Principles	33
2.4.2	Techniques	34
2.5	Document Planning	35
2.5.1	Principles	35
2.5.2	Techniques	36
2.6	Microplanning	37
2.6.1	Lexical Choice	37
2.6.2	Generating Referring Expressions	39
2.6.3	Aggregation	40
2.7	Surface Realisation	41
2.7.1	Principles	41
2.7.2	Techniques	42
2.8	Template NLG	43
2.8.1	Principles	44
2.8.2	Techniques	45
2.9	Further Reading and Resources	46
3	Machine Learning and Neural NLG	49
3.1	Examples	50
3.1.1	Very Simple Trained Model: <i>a</i> vs. <i>an</i>	50
3.1.2	Fine-Tuned Neural Model: Facebook Weather Dialogues	52
3.1.3	Prompted Model: Using ChatGPT to Generate Weather Forecasts	53
3.2	Machine Learning Models for NLG	55
3.2.1	Classifiers	55
3.2.2	N-Gram Language Models	56
3.2.3	Early Neural Models	57
3.2.4	Transformers and Foundation Models	57
3.2.5	Instruction Tuning and RLHF	58
3.2.6	End-to-end vs. Modular Architectures	60
3.3	Training Data	60
3.3.1	Data Sources	61
3.3.2	Data Set Criteria	62
3.3.3	Impact of Training Data on Prompted Models	63
3.3.4	Synthetic Data and Data Augmentation	64
3.3.5	Including Examples in Prompts	65
3.4	Issues	65
3.4.1	Domain Shift	66

3.4.2	Question Answering	67
3.4.3	Auditability and Controllability	68
3.4.4	Legal and Regulatory Issues	68
3.5	Further Reading and Resources	69
4	Requirements	71
4.1	Quality Criteria: Texts	72
4.1.1	Readability and Fluency	73
4.1.2	Accuracy	74
4.1.3	Content	76
4.1.4	Utility	76
4.2	Quality Criteria: Systems	77
4.2.1	Non-functional Requirements	77
4.2.2	Consistency and Variation	78
4.2.3	Average vs. Worst Case	79
4.3	Workflow	80
4.3.1	Fully Automatic NLG	80
4.3.2	Human Checking and Editing	80
4.3.3	Creating Drafts for Human Writers	81
4.4	Text and Graphics	82
4.4.1	Decision Support	83
4.4.2	Other Use Cases	84
4.4.3	Combining Text and Graphics	84
4.5	Requirements Acquisition	85
4.5.1	User Studies	85
4.5.2	Manual Corpus Analysis	88
4.5.3	Stakeholders	91
4.6	Further Reading	92
5	Evaluation	93
5.1	Example: Smoking Cessation (Impact Evaluation)	94
5.2	Fundamentals	96
5.2.1	Stakeholder Perspective	96
5.2.2	Hypothesis Testing	97
5.2.3	Statistical Hypothesis Testing	98
5.2.4	Experimental Design, Execution, Reporting, and Follow-Up	99
5.2.5	Research Questions	101
5.2.6	Replication	101
5.2.7	Ecological Validity: Artificial Versus Real-World Context ...	103
5.2.8	Test Data: Representative, Different from Training Data	104
5.3	Human Evaluation	105
5.3.1	Types of Human Evaluation	105
5.3.2	Experimental Design	110
5.3.3	Issues in Human Evaluation	118
5.4	Automatic Evaluation	121

5.4.1	Types of Automatic Evaluation	122
5.4.2	Experimental Design	124
5.4.3	Examples of Metrics	128
5.4.4	Validation of Metrics	130
5.5	Impact Evaluation	134
5.5.1	Comparison Between Users: Randomised Controlled Trials and A/B Testing	134
5.5.2	Historical Comparison	134
5.5.3	Challenges	135
5.6	Commercial Evaluation	136
5.6.1	Costs	136
5.6.2	Benefits	137
5.6.3	Risks	137
5.6.4	Return on Investment (ROI)	138
5.7	Ten Tips on Evaluating NLG	139
5.8	Further Reading	140
6	Safety, Testing, and Maintenance	143
6.1	Safety	143
6.1.1	Safety Concerns in NLG	144
6.1.2	Approaches to Addressing Safety Concerns	148
6.2	Software Testing of NLG Systems	151
6.2.1	Testing Systems with Variable Outputs	152
6.3	Maintenance	152
6.3.1	Changes in Domain and User Needs	154
6.3.2	New Users and Use Cases	155
6.3.3	Changes in Models	156
6.4	Further Reading and Resources	157
7	Applications	159
7.1	Key Attributes of Successful NLG Applications	159
7.1.1	Volume and Scalability	159
7.1.2	Data Availability	160
7.1.3	Accuracy	161
7.1.4	Maintainability and Adaptability	161
7.1.5	Acceptability and Trust	163
7.1.6	Conforming to Genre and Sublanguage	164
7.2	Journalism	164
7.2.1	Example: BBC Election Reporter	165
7.2.2	Types of News	166
7.2.3	Fake News	168
7.3	Business Intelligence	169
7.3.1	Example: Covid Reporter	169
7.4	Summarisation	170
7.4.1	Example: Summarising Emails with Google Bard	171
7.5	Medical Applications	172

7.5.1	Use Case: Reporting	173
7.5.2	Use Case: Patient Information and Behaviour Change	175
7.5.3	Use Case: Clinical Decision Support	176
7.5.4	Medical Business Intelligence Use Cases	177
7.5.5	Safety	178
7.6	Further Reading	179
References		181
Index		197