Storytelling from Structured Data and Knowledge Graphs
An NLG Perspective

Abhijit Mishra Anirban Laha Karthik Sankaranarayanan Parag Jain Saravanan Krishnan
IBM Research
{abhijimi , anirlaha , kartsank , pajain34 , sarkris5}@in.ibm.com

1 Goal of the Tutorial

In this tutorial, we wish to cover the foundational, methodological, and system development aspects of translating structured data (such as data in tabular form) and knowledge bases (such as knowledge graphs) into natural language. The attendees of the tutorial will be able to take away from this tutorial, (1) the basic ideas around how modern NLP and NLG techniques could be applied to describe and summarize textual data in format that is non-linguistic in nature or has some structure, and (2) a few interesting open-ended questions, which could lead to significant research contributions in future.

The tutorial aims to convey challenges and nuances in translation structured data into natural language forms, data representation techniques, and domain adaptable solutions. Various solutions, starting from traditional rule based/heuristic driven and modern data-driven and ultra-modern deep-neural style architectures will be discussed, and will be followed by a brief discussion on evaluation and quality estimation. A significant portion of the tutorial will be dedicated towards unsupervised, scalable, and adaptable solutions, given that systems for such an important task will never naturally enjoy sustainable large scale domain independent labeled (parallel) data.

2 Tutorial Overview

Natural Language Generation (NLG) has undergone significant advancement in the recent past, and various NLG systems are being used for either *data-to-text* tasks (*e.g.* generating financial reports from tables, generating weather reports) or *text-to-text* tasks (*e.g.* summarizing news reports, text-style transfer).

Structured data and knowledge bases or knowledge graphs are a key machine representation mechanism used in a wide variety of domains to capture domain-specific knowledge. For example, 1) the financial performance of companies and industries in financial domain, or 2) information about chemical composition of drugs, patient records, *etc.* in healthcare domain, or 3) inventory records of products and their features in retail, are all captured with domain-specific KGs/KBs. For AI driven interaction applications, often times it is important to communicate the content being represented in such knowledge bases in the form of natural language (such as English). Take an example in question-answering setting in Financial domain where a question:

"*How did XYZ corp. perform compared to its competitors in North America in last 2 quarters?*" would query a DB/KG and retrieves a result set table containing the relevant financial performance numbers about revenues, profit margin, competitors, technology segments, quarterly breakdown, etc.. However, it is not just sufficient for an AI system to simply display such a table of numbers, but rather, go one step further and explain in plain natural language the key message that addresses the user's question, for example, by saying,

"*In the N.A. region, XYZ Corp's revenues in the Cloud segment increased by 11% to $8.9B in the last 2 quarters as compared to its key competitor Microsoft. However, in the Analytics segment its revenues declined by 3% while Microsoft revenues grew by 4% and that of other smaller players in Analytics increased much more (around 8%).*"

Another important use-case is *story-telling* from data such as report generation – for example in weather domain (localized weather reports), finance (company performance reports) or healthcare (patient reports).

Motivated by above, this first-of-its kind tutorial intends to provide the conceptual underpinnings of the natural language generation (NLG) from a variety of structured representations. We will discuss various NLG paradigms ranging from heuristics to the modern data-driven techniques that include end-to-end neural architectures. A brief overview of evaluation methods and output quality estimation techniques will also be provided.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, pages 43–48
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

3 Type of the tutorial

Cutting-edge : We believe this topic picked up steam in the recent years given the deluge of papers regarding *data-to-text*. To the best of our knowledge, this topic has not been covered in any ACL/EMNLP-IJCNLP/NAACL tutorial.

4 Content of the Tutorial

We plan to organize a three-hour tutorial based on the following content. We will make efforts to make the tutorial interactive by having quizzes at regular intervals and also we hope to accommodate questions in between:

4.1 PART-I

1. **Introduction to NLG from Structured data and Knowledge Bases (20 mins)**
   - Data-to-text and text-to-text paradigms
   - Motivation: Why is this problem is important
   - Challenges in structured data translation: Why known text-to-text methods can not be applied to this problem?
   - Roadmap of the tutorial
2. **Heuristic Driven Methods (20 mins)**
   - Rule-based approaches
   - Template-based approaches
   - Current industry solutions
   - Shortcomings of this paradigm
3. **Statistical and Neural Methods (30 mins)**
   - Probabilistic Generation Models
   - Context-free Grammar based Approaches
   - Three-phase Approach : Planning, Selection and Surface Realization
   - End-to-end Encoder Decoder Paradigm
   - seq2seq approaches with attention
4. **Evaluation Methods for NLG (10 mins)**
   - N-gram based methods : BLEU, ROUGE
   - Document similarity based methods
   - Task-specific evaluation
   - Human evaluation metrics

4.2 PART-II

1. **Hybrid Methods - More adaptable (20 mins)**
   - Structured data input formats
   - Canonicalization
   - Simple Language Generation
   - Ranking of simple sentences
   - Sentence Compounding
   - Coreference Replacement
2. **Role of Semantics and Pragmatics (15 mins)**
   - Role of Knowledge Graphs
   - Domain-specific ontologies
   - Reasoning and Inference in Generation
3. **Open Problems and Future Directions (20 mins)**
   - Structure-aware Generation
   - Theme/Topic based Generation
   - Argumentative Text Generation
   - Controllable Text Generation
   - Creative Text Generation
4. **Conclusion and Closing Remarks (15 mins)**

Below we provide a bit more details about each of the above proposed sections to be covered in this tutorial.

Introduction to NLG from Structured data and Knowledge Bases: According to (Nema et al., 2018), the approaches for NLG range from (i) rule based approaches (ii) modular statistical approaches which divide the process into three phases (planning, selection and surface realization) and use data driven approaches for one or more of these phases (iii) hybrid approaches which rely on a combination of handcrafted rules and corpus statistics and (iv) the more recent neural network based models. Recent availability of large-scale parallel datasets like WIKIBIO (Lebret et al., 2016), WEBNLG (Gardent et al., 2017) have been like a catalyst for the recent research in NLG from structured data using data-driven neural models. However, modern NLG still faces challenges in various phases of content selection, surface realization and evaluation, as pointed out by Wiseman et al. (2017).

Heuristic Driven Methods: This paradigm was followed by early research in NLP and NLG (*e.g.*, (Dale et al., 2003; Reiter et al., 2005; Green, 2006; Galanis and Androutsopoulos, 2007; Turner et al., 2010)). They range from rule-based techniques to template-based techniques. Often these approaches involve choosing the right set of rules or retrieving the appropriate template for the generation task. Many popular industry solutions like Arria NLG[1] and Automated Insights[2] also follow this approach. As evident, there can only be a limited number of cases which can be handled by rules or that templates can cover. Hence, this approaches are not scalable or adaptable, paving the way for statistical approaches.

Statistical and Neural Methods: These approaches were formulated to alleviate some limitations of the earlier approaches. Some notable approaches are based on probabilistic language generation process (Angeli et al., 2010), context-free grammar based generation (Konstas and Lapata, 2012) and others (Barzilay and Lapata, 2005; Belz, 2008; Kim and Mooney, 2010). They popularized the three-phase paradigm by breaking the problem into three phases, namely, content planning, content selection and surface realization. The more recent neural approaches following the encoder-decoder paradigm, however, have tried to circumvent the three-phase approach by using a single-phase end-to-end architecture. This was mainly popularized by the advent of attention mechanism for seq2seq (Bahdanau et al., 2014), later followed by many (Mei et al., 2016; Lebret et al., 2016; Nema et al., 2018; Jain et al., 2018; Bao et al., 2018). However, these approaches are data-hungry and perform miserably on datasets from *unseen* domains (Gardent et al., 2017). Realizing this, some of the very recent works in data-to-text generation such as Wiseman et al. (2018) have focused on learning templates from corpora for neural NLG.

Evaluation Methods for NLG: Alongside discussion of methods for automatic generation of natural language, it is much needed to acquaint the participants about automatic evaluation metrics like BLEU(Papineni et al., 2002), ROUGE(Ganesan, 2018), METEOR(Banerjee and Lavie, 2005), among many others. Often, a different kind of evaluation is needed to measure the semantic relatedness which the above N-gram overlap based metrics may not always capture. In addition, for various NLG tasks, specialized metrics have been proposed like FleschKincaid for readability and SARI (Xu et al., 2016) for text simplification. However, the automatic metrics are not always enough to capture nuances like fluency, adequacy, coherence and correctness, which many NLG systems fallback on humans for evaluation.

Hybrid Methods: Some earlier approaches like (Langkilde and Knight, 1998; Soricut and Marcu, 2006; Mairesse and Walker, 2011) try to follow a combination of rules and corpus statistics to overcome the above shortcomings. In this portion of the tutorial, we are going to present a hybrid modular approach developed by us which can be broken down into three simple steps: (1) Canonicalization, (2) Simple Language Generation, and (3) Discourse synthesis and Language Enrichment. This has been developed in a domain-agnostic way without the need for any parallel corpora to train. This is not very data dependent and adaptable to various unseen domains as the generation steps are mostly restricted to linguistic aspects. We believe this is how the data-to-text generation research should progress.

Role of Semantics and Pragmatics: In this section we point out shortcomings of the above approaches which consider only surface-level characteristics for generation. Through this we motivate the necessity of knowledge graphs and domain-specific ontologies to understand the concepts present in structured data and assist the generation step through a deeper understanding. In this section, we will present a unification of literature from knowledge graphs area, like entity resolution, relation canonicalization, *etc.*, KG embeddings as well as heuristics which encode domain-specific pragmatics coupled with NLG to infer and produce higher-level and more complex natural language discourse.

Open Problems and Future Directions: This part will focus on various aspects of natural language generation which are far from being realized. The presenters will get highly creative and also borrow connections from some recent trends (Jain et al., 2017; Munigala et al., 2018; Hu et al., 2017; Jain et al., 2019) in NLG literature to formulate future directions for automatic text generation. The goal of this section is not only to motivate and

[1] `https://www.arria.com/`

[2] `https://automatedinsights.com/`

convey open research problems, but mainly to start a discussion paving the way for newer problems in the area.

Conclusion and Closing Remarks: We close with discussions about all approaches and some practical (as well as funny) observations for practical NLG realizations.

5 URLs

Slides: https://drive.google.com/open?id=1HaGCNc6n_sjyGLdaGzAVPvAeT0ZhhL3Q
Website: https://sites.google.com/view/acl-19-nlg

6 Breadth

This tutorial has more than 60% material which are not research outputs of the presenters. Thus majority of the material covered is discussion of the work done by other researchers.

7 Prerequisite Knowledge

We would like to ensure that the tutorial is self-contained. We do not assume any specific expertise from the audience. However, general awareness about Natural Language Processing and Machine Learning, and Deep Learning methods (such as Recurrent Neural Network, and Sequence-to-Sequence models) will be helpful.

8 Presenter Details

Abhijit Mishra
(https://abhijitmishra.github.io)
is currently a part of IBM Research AI, Bangalore, India, serving as Research Scientist in the division of AI-Tech. He is involved in multiple projects based on Natural Language Generation (NLG), viz. (1) Controllable Text Transformation (2) Structured Data Summarization, and (3) Devising evaluation metrics for quality estimation of NLG Output. Prior to joining IBM Research, he was a Ph.D. student in the Department of Computer Science and Engineering, Indian Institute of Technology Bombay (graduated in 2017). Since 2013, Abhijit's works have been consistently getting published in the proceedings of prestigious NLP/AI conferences such as ACL, AAAI, and WWW. He has also given multiple talks in Cognitive NLP, and Natural Language Understanding and Generation. The full list of his publications and talks are available in his website.

Anirban Laha
(https://anirbanl.github.io/)
is currently associated with the AI Tech group at IBM Research AI - India. He is interested in applications of machine learning/deep learning in natural language processing. He has been working in natural language generation (NLG) project in IBM for the last two years and has published papers on abstractive summarization both from unstructured and structured data in top conferences like NeurIPS, ACL and NAACL. At IBM, he has also worked on argumentation mining (IBM Project Debater[3]), which received news coverage worldwide recently because of a live machine vs human debate[4]. He was also briefly associated with machine learning for creativity project at IBM (SIGKDD workshop[5]), during which he has worked on story generation. Before joining IBM, he had spent some time as Applied Scientist in Microsoft and SDE at Amazon.com. He had received his MS degree from Indian Institute of Science (IISc), Bangalore. He had given talks on NLG, particularly NLG from structured data in multiple venues. The full list of his publications and talks are available in his website.

Karthik Sankaranarayanan
(http://bit.do/gscholar-karthik)
is a Senior Research Scientist and Research Manager working in the area of Artificial Intelligence at IBM's India Research Lab in Bangalore. He is currently leading research projects focused around Natural Language Generation (NLG), question-answering (QA), multimodal deep learning, and information retrieval from domain-specific knowledge graphs (NLQ) as part of IBM Watson. He has also managed efforts around argumentation mining (IBM Project Debater[3]), which received news coverage worldwide recently because of a live machine vs human debate[4]. He has published in flagship AI and knowledge management conferences and journals such as NeurIPS, CVPR, AAAI, IJCAI, ACL, NAACL, Machine Learning Journal, KDD, SIGMOD, VLDB, among others. He is an active PC member at several top academic conferences in AI. His innovations have resulted in more than 30 patents around applications of AI to industry problems. He is a Senior Member of IEEE. Before joining IBM Research in 2011, he obtained

[3] http://bit.do/ibm-project-debater
[4] http://bit.do/theverge-debater
[5] https://ml4creativity.mybluemix.net/

his PhD in Computer Science from The Ohio State University. Recently, he was the lead organizer of "Machine Learning for Creativity" workshop[5] at SIGKDD 2017, held at Halifax, Canada which was co-organized by IBM, Google Brain, Sony CSL. This workshop was attended by around 50 researchers from academia as well as industry and featured keynote talks by faculty from Harvard, MIT, amongst other notable researchers.

Parag Jain (https://parajain.github.io/) is currently working as a Research Engineer in IBM India Research Lab. At IBM he has been working in natural language generation (NLG) and has published papers on summarization from tabular data in top NLP conference like NAACL-HLT. He has also briefly worked on ontology driven dialog systems focusing on template based natural language generation from structured outputs. Recently, he has also published on Unsupervised Controllable Text Formalization in AAAI 2019. Parag completed his Masters in Computer Science from IIT Hyderabad in 2015. His M.Tech thesis was titled "Metric Learning for Clustering in Streaming Large-Scale Data". Prior to joining IBM, he has worked at Amazon.com as an SDE for a year. His website has all details about his publications.

Saravanan Krishnan (http://bit.do/linkedin-saravanan) is currently associated with the AI Tech group at IBM Research AI India since 2014. He is interested in natural language processing generation, mono and cross lingual information retrieval, information extraction, data mining and applications of machine learning. He has been working in natural language generation (NLG) project in IBM for the last two years focusing on research-oriented solutions for industrial deployments. Earlier at IBM, he was part of information retrieval group in IBM Project Debater[3], which received news coverage worldwide recently because of a live machine vs human debate[4]. Before joining IBM, he was at Microsoft Research India as Software Development Engineer for 6 years and at Anna University, Chennai as Project Associate for five years. He has published many papers in conferences (LREC, CIKM, ECIR, EACL) and journals (AJIT, LNCS) in the past 15 years of his research career. His LinkedIn profile has more details.

References

Gabor Angeli, Percy Liang, and Dan Klein. 2010. A simple domain-independent probabilistic approach to generation. In *EMNLP*, EMNLP '10.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. In *ICLR 2015*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72. Association for Computational Linguistics.

Junwei Bao, Duyu Tang, Nan Duan, Zhao Yan, Yuanhua Lv, Ming Zhou, and Tiejun Zhao. 2018. Table-to-text: Describing table region with natural language. *arXiv preprint arXiv:1805.11234*.

Regina Barzilay and Mirella Lapata. 2005. Collective content selection for concept-to-text generation. In *EMNLP*, HLT '05.

Anja Belz. 2008. Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models. *Nat. Lang. Eng.*, 14(4):431–455.

Robert Dale, Sabine Geldof, and Jean-Philippe Prost. 2003. Coral : Using natural language generation for navigational assistance. In *Twenty-Sixth Australasian Computer Science Conference (ACSC2003)*, volume 16 of *CRPIT*, pages 35–44, Adelaide, Australia. ACS.

Dimitrios Galanis and Ion Androutsopoulos. 2007. Generating multilingual descriptions from linguistically annotated owl ontologies: The naturalowl system. In *ENLG*, ENLG '07.

Kavita Ganesan. 2018. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. *arXiv preprint arXiv:1803.01937*.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for micro-planners. In *ACL*.

Nancy Green. 2006. Generation of biomedical arguments for lay readers. In *Proceedings of the Fourth International Natural Language Generation Conference*, INLG '06, pages 114–121, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1587–1596, International Convention Centre, Sydney, Australia. PMLR.

Parag Jain, Priyanka Agrawal, Abhijit Mishra, Mohak Sukhwani, Anirban Laha, and Karthik Sankaranarayanan. 2017. Story generation from sequence of independent short descriptions. *CoRR*, abs/1707.05501.

Parag Jain, Anirban Laha, Karthik Sankaranarayanan, Preksha Nema, Mitesh M. Khapra, and Shreyas Shetty. 2018. A mixed hierarchical attention based encoder-decoder approach for standard table summarization. In *NAACL-HLT*.

Parag Jain, Abhijit Mishra, Amar Prakash Azad, and Karthik Sankaranarayanan. 2019. Unsupervised controllable text formalization. In *AAAI*.

Joohyun Kim and Raymond J. Mooney. 2010. Generative alignment and semantic parsing for learning from ambiguous supervision. In *COLING*, COLING 10.

Ioannis Konstas and Mirella Lapata. 2012. Unsupervised concept-to-text generation with hypergraphs. In *NAACL-HLT*.

Irene Langkilde and Kevin Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *ACL*, ACL '98, pages 704–710, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *EMNLP*.

Franois Mairesse and Marilyn A. Walker. 2011. Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics*, 37(3):455–488.

Hongyuan Mei, Mohit Bansal, and Matthew R. Walter. 2016. What to talk about and how? selective generation using lstms with coarse-to-fine alignment. In *NAACL-HLT*.

Vitobha Munigala, Abhijit Mishra, Srikanth G. Tamilselvam, Shreya Khare, Riddhiman Dasgupta, and Anush Sankaran. 2018. Persuaide ! an adaptive persuasive text generation system for fashion domain. In *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, pages 335–342.

Preksha Nema, Shreyas Shetty, Parag Jain, Anirban Laha, Karthik Sankaranarayanan, and Mitesh M. Khapra. 2018. Generating descriptions from structured data using a bifocal attention mechanism and gated orthogonalization. In *NAACL-HLT*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Ehud Reiter, Somayajulu Sripada, Jim Hunter, Jin Yu, and Ian Davy. 2005. Choosing words in computer-generated weather forecasts. *Artif. Intell.*, 167(1-2):137–169.

Radu Soricut and Daniel Marcu. 2006. Stochastic language generation using widl-expressions and its application in machine translation and summarization. In *ACL*, ACL-44, pages 1105–1112, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ross Turner, Somayajulu Sripada, and Ehud Reiter. 2010. Generating approximate geographic descriptions. In *ENLG*, volume 5790.

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2017. Challenges in data-to-document generation. *arXiv preprint arXiv:1707.08052*.

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2018. Learning neural templates for text generation. *arXiv preprint arXiv:1808.10122*.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.